E.T.S.I.S. TELECOMUNICACIÓN

PROYECTO FIN DE CARRERA
PLAN 2000

TEMA:	BIG DATA: HADOOP
TÍTULO:	BIG DATA: ANALISIS Y ESTUDIO DE LA PLATAFORMA HADOOP
AUTOR:	NOELIA JIMÉNEZ BARQUÍN
TUTOR:	CARLOS CARRILLO SANCHEZ **Vº Bº.**
DEPARTAMENTO:	DIATEL
Miembros del Tribunal Calificador:	
PRESIDENTE:	LUIS ARRIERO ENCINAS
VOCAL:	CARLOS CARRILLO SANCHEZ
VOCAL SECRETARIO:	SARA LANA SERRANO
DIRECTOR:	
Fecha de lectura:	30-SEPTIEMBRE-2014
Calificación:	**El Secretario,**

RESUMEN DEL PROYECTO:

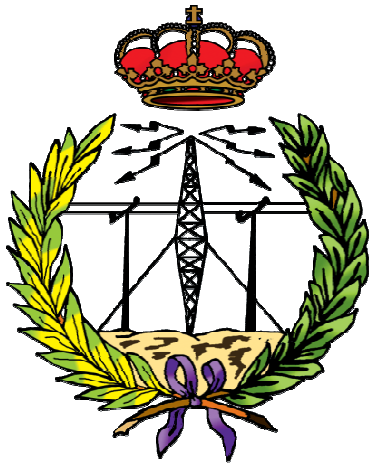
"El 90% de los datos de Internet se generaron en los últimos dos años", es el prefecto titular para explicar la necesidad imperiosa de una solución que permita el almacenamiento, gestión, visualización y manipulación de la ingente cantidad de datos que generamos en Internet.

Al estudio de esta solución le denominamos Big Data, que se encarga precisamente de ésto mediante plataformas de libre licencia como Hadoop.

En este PFC se tratará con el concepto de Big Data y se realizará un estudio sobre la plataforma Hadoop, sobre su arquitectura, tipos de datos, usos, sistemas de almacenamiento, su crecimiento y brevemente sobre su competencia.

UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Universitaria de Ingeniería Técnica de Telecomunicación



PROYECTO FIN DE CARRERA

BIG DATA: ANÁLISIS Y ESTUDIO DE LA PLATAFORMA HADOOP

NOELIA JIMÉNEZ BARQUÍN

SEPTIEMBRE 2014

AGRADECIMIENTOS

A Kira y Loewe por no llevar pilas,

a mi familia, porque sé que siempre estáis ahí

y a mi marido, porque sin ti nada sería posible.

I. ÍNDICE

II. ÍNDICE DE ILUSTRACIONES

III. RESUMEN

Desde el inicio de los tiempos el ser humano ha tenido la necesidad de comprender y analizar todo lo que nos rodea, para ello se ha valido de diferentes herramientas como las pinturas rupestres, la biblioteca de Alejandría, bastas colecciones de libros y actualmente una enorme cantidad de información informatizada. Todo esto siempre se ha almacenado, según la tecnología de la época lo permitía, con la esperanza de que fuera útil mediante su consulta y análisis.

En la actualidad continúa ocurriendo lo mismo. Hasta hace unos años se ha realizado el análisis de información manualmente o mediante bases de datos relacionales. Ahora ha llegado el momento de una nueva tecnología, Big Data, con la cual se puede realizar el análisis de extensas cantidades de datos de todo tipo en tiempos relativamente pequeños.

A lo largo de este libro, se estudiarán las características y ventajas de Big Data, además de realizar un estudio de la plataforma Hadoop. Esta es una plataforma basada en Java y puede realizar el análisis de grandes cantidades de datos de diferentes formatos y procedencias. Durante la lectura de estas páginas se irá dotando al lector de los conocimientos previos necesarios para su mejor comprensión, así como de ubicarle temporalmente en el desarrollo de este concepto, de su uso, las previsiones y la evolución y desarrollo que se prevé tenga en los próximos años.

IV. ABSTRACT

Since the beginning of time, human being was in need of understanding and analyzing everything around him. In order to do that, he used different media as cave paintings, Alexandria library, big amount of book collections and nowadays massive amount of computerized information. All this information was stored, depending on the age and technology capability, with the expectation of being useful though it consulting and analysis.

Nowadays they keep doing the same. In the last years, they have been processing the information manually or using relational databases. Now it is time for a new technology, Big Data, which is able to analyze huge amount of data in a, relatively, small time.

Along this book, characteristics and advantages of Big Data will be detailed, so as an introduction to Hadoop platform. This platform is based on Java and can perform the analysis of massive amount of data in different formats and coming from different sources. During this reading, the reader will be provided with the prior knowledge needed to it understanding, so as the temporal location, uses, forecast, evolution and growth in the next years.

1. INTRODUCCIÓN

Actualmente se calcula que existen 7.000 millones de móviles en el mundo, 500 millones de usuarios en Twiter, 1110 millones de cuentas de Facebook, a los que hay que añadir las cuentas de Instagram, de correo electrónico, etc. En 2012 eran 2.300 millones de internautas que se prevé lleguen hasta los 3.600 millones en 2017.

Todos los usuarios de Internet generan una gran cantidad de datos, a los que a simple vista, no se les da la importancia que tienen. Además hay que añadir información que se obtiene de manera indirecta mediante cookies, o del comportamiento de cada usuario, como acceder a periódicos o comercio online, dónde y cuándo, calcular la ruta al destino elegido para unos días festivos, hasta consultar la previsión meteorológica. Todo esto que para algunos carece de otro sentido, para otros puede esconder múltiples y variados propósitos, muy diferentes de los que se presumía.

Los datos obtenidos hablan de las personas, de sus gustos y de su personalidad. Con ellos grandes compañías generan perfiles a los que adaptar sus productos, y ofrecen al consumidor aquello que los "Big Data" dicen que quiere con campañas adaptadas y con productos pensados por y para ellos.

Además con Big Data se puede analizar el funcionamiento de aplicaciones software, detectar la probabilidad de futuras averías, mejorar la fabricación de dispositivos, etc. Todo ello mediante la instalación de sensores de captación de datos.

También se pueden generar aplicaciones que detectan epidemias, como por ejemplo brotes de gripe e incluso malaria, variaciones favorables en los precios de los comercios, localización de nuevos planetas y estrellas… y es que Big Data es aplicable tanto en astronomía, como genética, medicina, economía, sociedad, comercio y un sinfín de campos que se trataran más adelante.

En los orígenes del tratamiento de la información se utilizaban técnicas de minería de datos, para el análisis de datos basados en el uso de estadísticas y de bases de datos para descubrir patrones existentes en los datos de los que se deseaba extraer información.

Como introducción, Big Data es un conjunto de datos de distintas naturalezas que se extraen y almacenan para ser analizados por herramientas en busca de información y patrones ocultos.

Los conceptos Big Data y minería de datos pueden parecer similares porque ambas tecnologías persiguen el mismo objetivo sobre los datos. Para ello se basan en unas pautas fijas comunes, como son la recopilación de datos, su almacenamiento, su procesamiento y posterior análisis de los resultados, aunque estas fases sean claramente diferentes en cuanto a medios utilizados.

Pero existen diferencias. Una de ellas es que Big Data es un sistema que permite el procesamiento de grandes cantidades de datos en un tiempo mínimo y que estos puedan ser de distinta naturaleza. Además la minería de datos es más laboriosa y manual de lo que pretende ser Big Data y abarca únicamente pequeñas parcelas de la totalidad. Otra diferencia es que la minería de datos está limitada por la tecnología y el tipo de datos con los que tiene la capacidad de trabajar.

Por tanto se puede afirmar que la existencia de la minería de datos ha propiciado el auge de otras tecnologías que actualmente se encargan de la extracción y análisis de datos de formas más efectivas.

Gracias a Big Data y a las mejoras que ofrece, se pueden beneficiar todos. Los clientes, ya que al adaptarse los productos y servicios a sus gustos y necesidades, encuentran aquello que quieren, cuando, donde y como las necesitan. Y las empresas, toman mejores decisiones, generan mejores productos y aumentan su calidad en base a datos realistas obtenidos del estudio de los datos recopilados, siendo por tanto más competitivas.

En este trabajo se va a analizar también la herramienta Hadoop. Esta fue creada en 2006 por Dough Cutting con la finalidad de procesar datos a nivel web y trabajar conjuntamente con el motor de búsqueda denominado Nutch. En 2008, su autor pasó a formar parte del equipo de Yahoo! donde continúo con el desarrollo de Hadoop. Finalmente Apache Software Foundation desarrolló esta herramienta.

Actualmente esta plataforma, con licencia Open Source, es capaz de almacenar, procesar y analizar datos, trabajando con hasta 40000 máquinas y con cantidades del orden de petabytes. Su desarrollo sigue creciendo por la gran comunidad de desarrolladores y usuarios, que han hecho de Hadoop un ecosistema completo con variedad de herramientas auxiliares.

La plataforma permite el procesamiento de datos de diferentes naturalezas y estructuras, provenientes de los más distintos medios: audio, video, texto, etc. Esta procesa todos los datos de forma simultánea y distribuida ofreciendo como resultado la relación que poseen los diferentes datos.

Como curiosidad, Cutting nombro a su proyecto “Hadoop” en honor a un elefante de juguete, lo que se ha convertido en su imagen.

2. BIG DATA

Como se adelantó en la introducción, en este capítulo se va a tratar de responder a algunas preguntas claves sobre Big Data que faciliten su comprensión y su uso en la actualidad.

2.1. ¿QUÉ ES?

Big Data se puede definir como un nuevo método impulsado por grandes compañías como IBM, EMC, Oracle o Yahoo, que gestiona, almacena y analiza grandes cantidades de datos que escapan del manejo con las herramientas comunes hasta el momento. Es la adaptación del manejo de datos a los tiempos actuales, con nuevas herramientas y capacidades de procesamiento. En este momento, se puede afirmar que el 90% de los datos existentes han sido generados en los últimos años, ello desborda los métodos conocidos hasta el momento para analizar y procesar los datos. Pero la mejor forma de entender qué es, es entender su utilidad y como funciona. Para ello se han definido en este capítulo unas preguntas claves, cuyas respuestas se deben conocer para adentrarse y comprender el concepto de Big Data y las aplicaciones y empresas que trabajan con este método.

2.2. OBJETIVOS

Big Data pretende dar un giro radical a los métodos de análisis y con ello mejorar la productividad, la toma de decisiones, reducir los errores y evasiones fiscales, detectar oportunidades de mercado, ahorrar tiempo y dinero a empresas y ciudadanos. Pretende trabajar para conseguir mejores soluciones en investigaciones científicas y médicas, para describir patrones en enfermedades y un sinfín de propósitos más.

En un principio y con las herramientas y conocimientos necesarios se verá que es posible conseguir todo esto. De hecho ya son muchos los que vieron el potencial en el nuevo método. Empresas como UPS se benefician desde hace años de su potencial. Con su uso ha conseguido optimizar rutas en sus repartos, prever problemas mecánicos en sus vehículos y con ello ahorrar en combustible, tiempo y dinero. Ello ha sido posible mediante el análisis y la comprensión de los datos recibidos por la multitud de sensores de los que disponen sus vehículos y coordenadas de GPS que realizan el seguimiento de sus rutas de envío.

Se puede resumir por lo tanto, a grandes rasgos, que Big Data tiene como principales objetivos el recolectar y analizar los datos en busca de información relevante para mejorar distintas situaciones de la vida, ayudando a tomar mejores y más rápidas decisiones que con las herramientas convencionales.

2.3. ¿CUÁNDO?

Big Data es recomendable siempre, pero se hace del todo evidente cuando debido a la cantidad de datos y su naturaleza, su procesamiento se hace una tarea imposible. Y se puede decir que es siempre, porque aún cuando la denominación Big Data hace pensar en tamaños de datos enormes, no es del todo cierto que sea su única utilidad. El análisis de pequeñas cantidades de datos puede también ser útil para captar correlaciones que de otro modo se pasarían por alto. Es entonces cuando se puede afirmar que es recomendable prácticamente en cualquier situación, independiente del caso y del tamaño de datos que sean procesados, se les puede dar valores añadidos y generar nuevas estrategias y modelos de mercado.

Además se hace necesario el uso de Big Data cuando los datos esconden información que no es visible a simple vista. En muchos casos al aplicar algoritmos de correlación[1] a los datos se descubren patrones imposibles de captar de ningún otro modo. En algunos casos revelan información curiosa, como pasó en el supermercado Wal-Mart en Estados Unidos. Según la noticia publicada en noviembre de 2004 por "The New York Times"[2], al realizar el análisis de los datos recopilados, los responsables encontraron una relación entre los periodos en que había riesgo de tornado y el consumo de palomitas de maíz. Este hecho, teniendo en cuenta los productos que han sido vendidos en un supermercado y las previsiones de tornado, no es fácil de detectar a simple vista, pero de este modo el supermercado consiguió sumar ventas en dichas temporadas ofreciendo a sus clientes el producto de forma más accesible en mejores y más visibles lugares del comercio.

Hechos como el de Wal-Mart, pueden pasar desapercibidos si sólo se tiene en cuenta la naturaleza y la cantidad de datos que se posee. Aunque no lo parezca a simple vista o no se deduzca con métodos antiguos de procesamiento, pueden ser muchas las informaciones que oculten y pueden generar ganancias o mejoras en las compañías e incluso en los gobiernos.

La mayor ventaja del uso de herramientas basadas en Big Data se obtiene cuando la cantidad de datos a procesar es muy elevada, ya que a mayor cantidad de datos se obtendrán mejores y más afinados resultados, siempre y cuando su elección haya sido correcta y los datos sean íntegros y no corruptos.

De cualquier modo cuando la cantidad de datos a almacenar y procesar es del orden de los Terabytes se hace del todo necesario su análisis mediante técnicas de Big Data para poder extraer de ellos todo su potencial. Además es evidente la necesidad de Big Data cuando se requiere un procesamiento rápido y el valor de los datos va ligado al tiempo de procesamiento y análisis. Aunque no se debe negar que Big Data puede

[1] Puede encontrarse más información sobre correlación en el glosario.
[2] http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html?_r=0

llegar a ser menos preciso que una base de datos relacional, ésta permite la obtención de mucha más información para analizar tras el procesamiento.

La naturaleza de los datos también es algo a tener en cuenta. En el caso de que estos sean de una naturaleza variada (datos estructurados, semi-estructurados y no-estructurados) es necesario el uso de plataformas de Big Data para su análisis, debido a que el análisis de datos mediante bases de datos relacionales solo permite trabajar con datos estructurados. Por lo que para no perder información referente a las distintas naturalezas de los datos, se necesitan plataformas que permitan trabajar con el conjunto completo de la información.

2.4. CARACTERÍSTICAS Y VENTAJAS

Big Data posee cuatro características básicas que lo hacen distinto de cualquier otro método de análisis de datos: volumen, velocidad, variedad y veracidad.

Ilustración 1. Características Big Data

Al contar con el manejo de grandes volúmenes de datos como ventaja, Big Data abre la puerta a asuntos de gobierno, empresariales y políticos, que manejan datos de miles de millones de habitantes. Con su análisis se permiten ahorros significativos fiscales mejorando la eficiencia y detección de errores en temas económicos. Además Big Data se caracteriza por su velocidad en el procesamiento de datos, aunque no deja de ser una herramienta que trabaja en lotes. Y es que en ocasiones la obtención de un valor en un breve periodo de tiempo puede producir ganancias importantes. Como por ejemplo en el caso de las carreras de Fórmula 1, donde la obtención y análisis de los datos provenientes de los sensores de un coche en el momento de la carrera puede ser vital para conseguir subir al podio. Esto actualmente podría solventarse de una forma mucho más rápida haciendo uso de Hadoop y no de las tradicionales bases de datos relacionales.

La variedad es una ventaja que se hace evidente cuando los datos a tratar son de distintas naturalezas, desde emails hasta videos, fotos y comentarios en redes sociales introducidos por la mano humana hasta valores numéricos obtenidos por sensores. Con Big Data existe la posibilidad de analizar estos datos juntos y sin importar su naturaleza, para obtener información que con los antiguos métodos era imposible. Es decir, Big Data permite analizar juntos datos estructurados, semi-estructurados y no estructurados.

Por último, la veracidad, es la característica más difícil de mantener en Big Data. Ello supone tener en cuenta la volatilidad y validez de los datos. Implica utilizar datos útiles, que no están malogrados en el tiempo o que son poco fiables y manejarlos con una cautela máxima para no dar análisis erróneos que puedan truncar decisiones.

Otras ventaja, no menos importante, es que Big Data en la actualidad esta soportado por plataformas de código abierto como se estudiará en otros capítulos. Esto permite que se acerque a conocer su potencial gran cantidad de público independientemente de su poder adquisitivo. Aunque trabajar con esta metodología requiere de un mínimo equipamiento, también es cierto que se ha visto ligada a un significativo descenso del precio del almacenamiento y el aumento de sus capacidades. Además de surgir en un momento en que las máquinas poseen una buena capacidad de procesamiento, permitiendo así un alto rendimiento en el análisis y lo que hace aún más posible que Big Data sea una opción rentable y no requiera de un presupuesto privativo.

Como valor añadido y aunque no es una característica propia de Big Data, sino de las diferentes plataformas que hacen uso de este método, es importante mencionar que ofrece escalabilidad, flexibilidad y es un sistema distribuido. Todas ellas son ventajas importantes a la hora de sopesar el uso de esta tecnología. Big Data está pensado para crecer y para adaptarse a nuevas formas y volúmenes de datos.

2.5. REQUISITOS

Para un manejo fluido de Big Data se requieren grandes anchos de banda. Esto no supone un gran problema teniendo en cuenta que en 2012 se calculó en torno a 11,3Mbps de ancho de banda que se prevé se multiplique por 3,5 para el año 2017 alcanzando unas cifras de hasta 39Mbps. Por lo tanto, este requisito se presupone fácilmente superable teniendo en cuenta la rapidez con la que evolucionan las redes, el ancho de banda y la tecnología en general.

Un requisito más indispensable, es disponer de una cantidad razonable de datos para poder sacarle el mayor beneficio. Aunque esto no es necesario, se requiere de un mínimo para poder extraer toda la ventaja de su uso.

Por último, para su correcto funcionamiento se requieren sistemas con un alto rendimiento de procesado y dependiendo de la cantidad de datos con los que se trabaje se requiere de una cantidad variable de nodos de almacenamiento.

2.6. HASTA HOY

Los datos han existido siempre. Desde la primera pintura rupestre, hasta el último dato almacenado en la más moderna máquina, todos forman parte de la información y del deseo de conservarlos, pero no siempre se ha tenido la capacidad para analizarlos adecuadamente y extraer de ellos todo el potencial que tenían oculto.

Actualmente, son muchas las empresas que han recogido gran cantidad de datos sobre las relaciones que mantenían con los clientes o aquellos que provienen de sensores que generaban automáticamente información. Estos se analizaban sin tener en cuenta el valor oculto que escondía. Pero desde hace un tiempo, y antes de que se asentara el concepto Big Data, ya se podía extraer de ellos importante información, aunque de una forma mucho más laboriosa que como se hace ahora.

Hasta hace poco tiempo la única forma de tratar datos era mediante bases de datos relacionales[3]. Actualmente la cantidad de datos que se generan diariamente lo hace imposible. El análisis mediante bases de datos relaciones es un trabajo que requiere de intervención humana y se complica al procesar grandes cantidades de datos. Además este análisis no tiene en cuenta la totalidad de los datos por su tipo de estructura.

Con Big Datase produce un gran avance, es posible analizar cualquier dato independientemente de su naturaleza y con un esfuerzo mínimo. Al encargarse de analizar la información relevante, permite obtener el mayor beneficio en menor tiempo, de forma más fácil y teniendo en cuenta el conjunto completo de los datos sea cual sea su naturaleza.

Utilizar plataformas de Big Data automatiza el proceso en gran medida. Permite analizar datos que antes eran almacenados sin saber bien el destino que correrían. Con este nuevo enfoque se está dando una segunda oportunidad a la información y se están descubriendo nuevos usos que los datos escondían. Además las empresas en ocasiones son capaces de sacarles un provecho doble, mediante la venta de esta información a otras empresas que pueden hacer un uso provechoso de ella.

Con el uso de plataformas de Big Data en vez de bases de datos relacionales, se está permitiendo obtener una respuesta de los datos muy rápida, lo cual no quiere decir que la respuesta sea inmediata. El objetivo de Big Data no es analizar pequeños datos a una gran velocidad, sino analizar cantidades importantes de datos sin la necesidad de

[3] Información ampliada sobre bases de datos relacionales en el glosario

una respuesta inmediata, pero en un tiempo inferior al empleado en BBDD relacionales. Obteniendo las respuestas en un mínimo periodo de tiempo se le dan a los datos un valor añadido, con ellos se consigue hacer uso más adecuado y acertado en el contexto en el que han sido extraídos. Al contrario ocurre mediante el análisis con bases de datos relacionales, al tomar más tiempo en el procesamiento, hace que el resultado del análisis pueda haber perdido su valor.

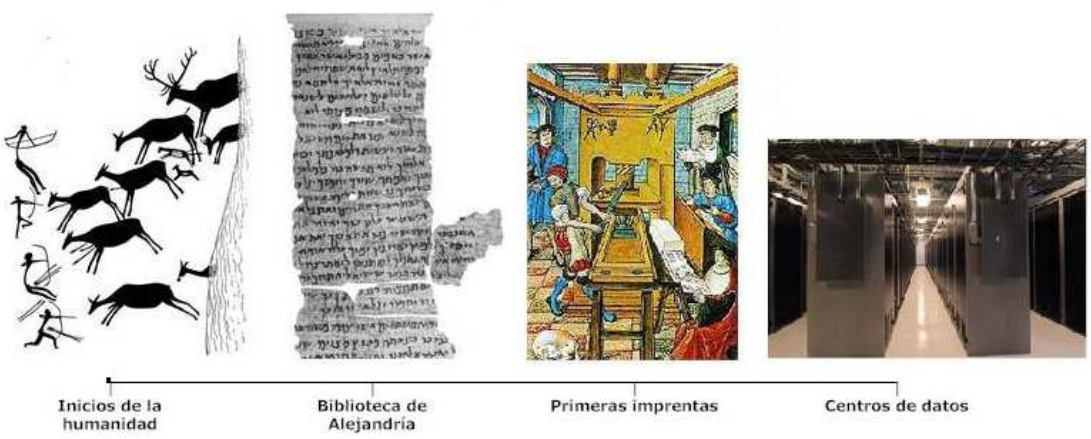


Ilustración 2. Desde los inicios de la humanidad se recopila, graba y almacena información

2.7. ¿DE DÓNDE VIENEN LOS DATOS?

Según el artículo *¿Qué es Big Data?* De la compañía IBM en el año 2012, se generan cerca de 2,5 quintillones de bytes diariamente en todo el mundo y prevé que el volumen de información crezca un 40% entre 2012 y 2020 llegando así a los 40 zeta bytes. Esta gran cantidad de datos son en su mayoría datos creados por Smartphones, sensores, tablets, televisiones, cámaras, medidores, comunicaciones entre máquinas (M2M), etc., utilizados en tan diversas temáticas como tráfico, comercio, medicina, meteorología, redes sociales, logística, investigaciones científicas y médicas, seguridad o educación.

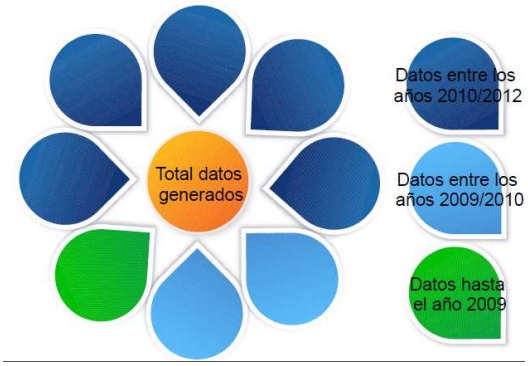


Ilustración 3. Volumen de información generada hasta el año 2012

Sin apenas ser conscientes de ello se generan diariamente una gran cantidad de datos con los Smartphone, cada imagen o comentario publicado en una red social, cada "me gusta" y cada acción en una red social suma el número de bytes que se están usando. También cuando se descargan aplicaciones móviles y cuando se conecta el GPS, wifi o bluetooth. Todo ello genera una cantidad importante de datos con los que realizar estudios sobre localización del usuario y demás. Existen un sinfín de datos útiles para generar un perfil sobre quién y donde está generando los datos y como y para qué está haciendo uso de ellos. Esto permite realizar mejoras en las ofertas y aplicaciones en el futuro.

Pero no solo se generan datos con los Smartphones. En los hogares simplemente con el uso de un teléfono fijo se generan datos que indican a las empresas los horarios favoritos para llamar, los destinos más frecuentes y la duración de las llamadas. Con esta información se generan planes de negocio adaptados a los gustos y necesidades del usuario.

En otras muchas otras ocasiones más, no se es consciente de que se están generando valiosos datos, como con el uso de gasolineras, carreteras, centros comerciales, cajeros automáticos y prácticamente en cualquier lugar imaginable.

También es cierto, que no todos los datos son de nueva creación. De igual modo que Google comenzó a realizar el escaneo de múltiples libros para añadirlos a su librería online, muchas empresas, compañías y gobiernos están computarizando multitud de datos que antes solo existían en un formato de papel. Esto permite extraer de ellos estadísticas e información que estaba oculta hasta el momento. Mediante el análisis de esos datos tomados en determinados años, sobre un mismo suceso, pueden extraerse patrones que alienten sobre sucesos futuros.

No solo se están computarizando datos que antes se tenían en otro formato, si no que a su vez se está fomentando la automatización de todos ellos, ya son pocos los ayuntamientos, bibliotecas y demás organismos que trabajen con datos tomados a mano y no con ordenadores, lectores RFC, lectores de códigos de barras, etc.

Para ser más conscientes de la magnitud de datos que se están tratando, a continuación se encuentran cifras[4] concretas sobre los grandes volúmenes de datos con los que se trabaja:

- Twiter: recibe más de 5700 twits por segundo.
- Whasapp: 64 millones diarios de mensajes en el mundo.
- Facebook: posee más de 1.15 millones de usuarios activos y más de 100TB de información al día almacenada.

[4] Cifras extraídas durante el año 2014.

- Visa: gestiona más de 173.000.000 operaciones de tarjetas diarias.
- Youtube: almacena 72 horas de video cada minuto.
- RFID[5]: la radiofrecuencia genera 1000 veces más información que los códigos de barras convencionales.
- UPS: realiza 39.5 millones de rastreos de envíos a sus clientes diariamente.
- Se crean más de 571 nuevas webs por minuto.

2.8. ALCANCE

Actualmente muchas empresas almacenan y manejan un considerable volumen de datos sobre sus clientes y su productividad de forma periódica, pero no todas las empresas han sabido sacar provecho a esta información o estaban preparadas para hacerlo.

Según el informe de la compañía Cisco, en 2013 el 60% de las compañías eran conscientes de la ventaja que los grandes datos les ofrecen en toma de decisiones y estrategia, pero sólo un 28% de ellas lo han puesto en práctica. Son muchas las que se plantean grandes mejoras a corto plazo en esta materia, en la mayoría de los casos los departamentos aseguran que Big Data es desde 2013 una parte importante de su trabajo y una prioridad estratégica.

Es importante tener en cuenta que con el uso de Big Data, además de obtener beneficios económicos, se puede hacer un uso eficiente de la información obtenida en distintos departamentos dentro de una empresa, como por ejemplo en recursos humanos, contabilidad, producción y demás. Además se puede hacer uso de Big Data para usos generales sobre toda la población, mediante estudios en hospitales, ayuntamientos y otros.

Un ejemplo interesante es FluThrends, herramienta generada por Google junto con el organismo de Control y prevención de enfermedades de Estados Unidos. Esta herramienta analiza geográficamente las búsquedas realizadas en el buscador Google referentes a la gripe y sus síntomas, a partir de ello se puede obtener una previsión de si existe una epidemia de gripe en el área geográfica estudiada. Para ello tiene en cuenta factores como la densidad de población, los casos reales de enfermedad los años anteriores, antiguos registros sanitarios, etc.

[5] Radio Frecuency IDentification: sistema de almacenamiento de datos convencionalmente en pequeñas etiquetas que contienen una antena para recibir y responder información solicitada por un emisor-receptor RFID; no requieren de alimentación eléctrica y pueden ser colocadas en productos, animales y demás.

2.9. DESVENTAJAS

La principal desventaja que presenta el uso de datos es la seguridad. Big Data se ve afectada por una gran cantidad de leyes que hacen que el trabajo con datos sea un tema realmente delicado. Existen casos en que se ha conseguido destapar la identidad de una persona sin permiso, mediante la correlación de datos personales sujetos a diferentes leyes de protección y datos obtenidos de sus compras, aficiones, localización geográfica, etc.

Como menciona el periódico ABC[6] en uno de sus artículos de Abril de 2014, hay quien ha conseguido evitar, no sin grandes esfuerzos y grandes precauciones, que sus datos sean tratados mediante Big Data. El caso concreto es una mujer que oculto a Big Data su embarazo, procurando que la información a través suyo y de sus familiares no se filtrara a la red a través de mensajes de Facebook, compras en Amazon y demás.

Debido a estos casos en los que sin quererlo puede ponerse al descubierto la identidad de una persona, se hace especialmente importante trabajar con una seguridad que no comprometa la identidad de sus dueños. Aunque en su mayoría el ser humano sólo considera información personal aquella como el número de teléfono, nombres, apellidos, DNI y poco más, son muchos otros datos los que pueden considerarse personales. Un ejemplo es la huella que se deja al acceder a cierta aplicación, al visitar una página web y demás. En esa huella puede encontrarse la dirección IP del usuario, la localización, horario, tipo de navegador o sistema operativo que se ha utilizado junto con otra información, que cruzada adecuadamente puede descubrir la identidad del usuario sin esfuerzo.

Es por ello, que la gran desventaja de la manipulación y análisis de grandes cantidades de datos procedentes de distintas fuentes, es el derecho a la intimidad, la seguridad y la propiedad intelectual. Ya son muchas las compañías que como IBM con “Guardium Inphosphere”[7] permiten la detección y corrección de posibles datos que ponen en juego la seguridad.

La otra gran desventaja es el presupuesto en el equipo tecnológico y la formación de los trabajadores. Aunque muchas empresas han comenzado a incluir en los presupuestos a partir del 2013 y 2014 un apartado significativo para Big Data, son también muchas las que aún no lo contemplan.

Las empresas han de realizar un gran esfuerzo en invertir en Big Data para aumentar ventas y tomar mejor decisiones. Ello también implica tener que realizar modificaciones en la infraestructura existente, preparar sus redes para un aumento

[6] http://www.abc.es/tecnologia/redes/20140430/abci-ocultar-embarazo-redes-data-201404292125.html
[7] Más información sobre Guardium Inphosphere en el glosario

significativo de tráfico, mejorar la seguridad en el sistema, adaptar los equipos a las nuevas capacidades y formar al personal para su utilización.

Al tratarse Big Data de un entorno no solo útil en un departamento dentro de una empresa, se considera importante formar a diferente personal, de modo que todos ellos al margen de si se trata de departamentos técnicos, jurídicos o demás puedan beneficiarse de la nueva tecnología. Cuando Big Data esta asimilado dentro de los distintos departamentos, no sólo los departamentos de marketing e informática se ven beneficiados, sino que lo será el conjunto de la empresa.

Esta formación del personal tiene una gran importancia dentro del planteamiento de implantar Big Data. Como se verá, el análisis erróneo o deficiente de los datos obtenidos en la fase de procesamiento puede hacer perder grandes oportunidades u obtener resultados erróneos a la vez de perder tiempo y dinero.

2.10. ¿QUÉ TIPO DE DATOS?

El tipo de datos con los que puede trabajar Big Data, concretamente la plataforma Hadoop, es precisamente una de sus principales ventajas y en gran parte lo que lo hace diferente de otras técnicas de manipulación y procesamiento de datos.

A diferencia de los métodos habituales de análisis de datos, éste permite el trabajo conjunto con datos estructurados, semi-estructurados y no estructurados.

Ilustración 4. Tipos de datos admitidos en Big Data

El trabajo con las distintas naturalezas de datos elimina una de las barreras que antes se podía encontrar al analizar la información usando otros métodos. Mediante Big Data es posible analizar con facilidad datos estructurados, al igual que ya era posible anteriormente. Estos son datos provenientes de cadenas "strings" como números de DNI, emails, nombres, direcciones, fechas, teléfonos, coordenadas y números en general. Este tipo de datos supone únicamente un 20% de los datos que genera una empresa, además lo son cada vez en menos proporción, debido al rápido crecimiento e implantación de nuevas tecnologías basadas en otros tipos de datos, como pueden fotos, videos y demás. Sólo con este dato ya se es consciente de la gran limitación que existe al poder realizar análisis de datos solo mediante bases de datos relacionales.

Ilustración 5. Datos de diferentes procedencias pueden ser procesados por Big Data

En la actualidad y muy probablemente en el futuro, Big Data seguirá trabajando con datos estructurados en gran medida ya que estos datos generalmente proceden de máquinas y sensores. Además seguirá trabajando con otros tipos, como datos semi-estructurados y no estructurados. Para hacerse una idea, se entiende como datos semi-estructurados aquellos que siguen por ejemplo el formato XML[8] y se entiende por datos no estructurados las imágenes, videos, datos concretos de una máquina o aplicación en un formato no estructurado.

[8] eXtensible Markup Language, lenguaje de marcas desarrollado por World Wide Web Consortium para ser utilizado en bases de datos, hojas de cálculo, editores de texto, etc, permitiendo almacenar datos de forma legible.

Es precisamente lo que hace que esta forma de análisis sea distinta a todas. Permite manejar imágenes, videos provenientes de Youtube, cámaras de vigilancia, dispositivos de video médico y científico, texto sin patrón recogido de twits, comentarios en Facebook, páginas web y emails, imágenes provenientes de investigaciones espaciales, imágenes subidas a internet y demás. Por último cualquier tipo de dato proveniente prácticamente de cualquier dispositivo independientemente de su formato: pdf, SMS, audio, etc.

2.11. ¿CÓMO?

El funcionamiento general de Big Data es independientemente de la plataforma en la que se trabaje. Se divide en cuatro pasos fundamentales:

- Captación de datos
- Almacenamiento de los datos
- Procesamiento de los datos (mediante plataformas que realizan correlación de datos)
- Análisis y adaptación de los resultados

En la fase de captación de datos, las empresas y organizaciones realizan una recolección de datos a través de clientes, sensores, datos internos y distintos dispositivos dependiendo de su finalidad. Todos estos datos son y han sido almacenados durante largos periodos de tiempo, en la mayoría de los casos para comprender un mayor número de casos posibles y así obtener mejores resultados más ajustados a la realidad. Ocurre que cuanto mayor es la muestra de datos que se toman, mejores resultados se pueden obtener, siempre que también la elección de los parámetros y el procesamiento hayan sido a su vez eficaces. Una vez obtenida y almacenada toda esta información es procesada. Para ello se requiere conocer con la mayor precisión el fin que se desea obtener. Como se verá a continuación, de nada sirve procesar una gran cantidad de datos si no han sido previamente seleccionados correctamente.

Por tanto, hay que dar una especial importancia a la captación de los datos. Para obtener resultados fiables, además de obtener el mayor número de datos posibles, con la mayor veracidad, es importante saber si se están eligiendo todo el rango o tipo de datos correctos.

En el caso de la herramienta antes mencionada Flu Trends de Google, se ha visto que los datos de los que parte, son el número de búsquedas de términos relacionados con la gripe en un periodo y localización determinados. Pero eso no es todo, también hace uso de datos como estadísticas de gripe en años anteriores en la misma situación geográfica y en el mismo periodo, antiguos registros sanitarios, número de casos reales

y uso del buscador Google en la zona de estudio. Este último parámetro es el que puede marcar una diferencia muy significativa. Para ser conscientes de la magnitud de un dato erróneamente seleccionado, puede imaginarse el caso en que se hace un estudio sobre la gripe mediante Flu Trends en la ciudad de California, donde la mayoría de la población no hace uso del buscador Google. Al obtener los resultados se puede comprobar que no son nada acertados. En muy pocos casos se obtendría como resultado que se está produciendo una epidemia de gripe, puesto que al no tener en cuenta la variable uso del buscador Google, se ha supuesto que la población hacia uso de este motor de búsqueda y realmente están haciendo sus búsquedas en otro buscador. Este es el inconveniente de no hacer una buena elección de los datos con los que se va a trabajar.

Este procesamiento de datos es la aplicación de una serie de normas estadísticas, correlaciones y teorías que encuentran patrones y acciones predecibles en el conjunto de datos que se procesan. Aunque resulte una terminología más familiar para cualquier persona especializada en el campo de la estadística, en el procesamiento de los datos se trata de encontrar siempre una correlación. Esto implica que dos o más variables estén correladas, por tanto, el cambio en uno o varios datos provocan un cambio en otro u otros datos.

Con todo esto, se obtienen gran variedad de patrones. Algunos de ellos permitirán predecir acciones y hechos que de otro modo pasarían desapercibidos y que hubieran sido imposibles de descubrir por el tipo de información o la imposibilidad de procesarla. Otros patrones serán únicamente ruido, no tendrán ningún valor, serán meras coincidencias. En algunos casos, aunque se dé la correlación de datos y la identificación de patrones, estos no implican una información valida. Como se estudiará más adelante en los errores y problemas de Big Data, ocurre que en ocasiones una correlación no justifica la existencia de una relación causa y efecto, por ello ha de ser analizada por personal entendido que descarte estos falsos positivos. Esto hace que una de las etapas más importantes de Big Data, sea contar con un personal que valore la información obtenida del procesamiento de los datos. Este personal será el que logre captar cual o cuales son realmente los patrones importantes que se puedan aplicar para su beneficio. En este aspecto, Big Data está muy estrechamente unido al campo de la estadística. Requiere de personal con conocimientos estadísticos, que analice la información obtenida del análisis de los datos y que deseche la información basura, correlaciones casuales y ruido. Converge por lo tanto en la gran importancia de un personal que tenga conocimientos para interpretar la información y obtener conclusiones válidas.

2.12. APLICACIÓN

El campo de aplicación de Big Data es amplio y variado. Resulta imposible mencionar todas las empresas y organizaciones que a diario hacen uso de Big Data, pero se pueden mencionar las más conocidas.

- Hamlet: más tarde conocida como Farecast y comprada por Microsoft a su creador Oren Etzioni, un profesor de Inteligencia Artificial en Seattle. Tras su propia experiencia decidió crear un algoritmo para predecir los precios de billetes de avión basándose en las variables, tipo de avión, distancia, época del año, horario, precio del combustible, etc.

- Sloan Digital Sky Survey (SDSS): proyecto que capta imágenes espaciales a través de un telescopio para diferentes objetivos, entre ellos la difusión de las imágenes a través de la Nasa y Google Sky. Hace uso de Big Data para su almacenamiento, distribución y procesamiento. En una sola noche el telescopio puede almacenar hasta 200 gigabytes de datos, desde el año 2000 ya ha almacenado 140 terabytes y se prevé que a partir del año 2016 se capten 140 terabytes de datos cada cinco días con su telescopio predecesor, el gran telescopio de rastreo sinóptico de investigación.

- Paypal: desde hace bastante tiempo Paypal ha recogido y almacenado datos sobre las transferencias y pagos que se realizan los clientes a través de su plataforma. Datos como el navegador desde el que se realiza el pago, sistema operativo del equipo, horas, direcciones IP, localización... todo ello para refinar la forma de detección de fraude. Con Big Data se hace posible analizar la gran cantidad de datos que reciben al minuto desde una infinidad de localizaciones, países, monedas... lo que antes era del todo imposible. La plataforma que utiliza actualmente es Hadoop.

- Universidad de Northwestern: ha desarrollado un método basado en Big Data para la predicción de tormentas y huracanes teniendo en cuenta los datos recogidos durante años sobre meteorología.

- Hospital La Fe de Valencia: a través de la compañía SAS el hospital La Fe ha implantado un sistema de análisis avanzado de datos mediante el cual mantener un control de los enfermos crónicos. Además obtienen una información detallada y completa para mejoras presentes y futuras en los tratamientos de dichos pacientes.

- Ebay: el centro de subastas online requiere procesar diariamente cerca de 50 terabytes y 100 millones de clientes de forma efectiva para obtener el mayor rendimiento y ganancia con sus ventas. Por ello apoya gran parte de sus

datos sobre una plataforma Hadoop formada por 20.000 nodos de 80 petabytes de capacidad. Con ellos tiene la capacidad de administrar y procesar de forma eficiente la información para fijar precios, realizar subastas, precios de envíos y demás información.

- AXA Seguros e inversiones: analiza datos con el fin de detectar fraudes, blanqueo de dinero y operaciones ilegales.
- ING Direct: recopila y analiza información sobre clientes para ofrecerles productos adaptados a las necesidades y tomar mejores decisiones en el negocio.
- Banco SABADELL: a través de las redes sociales Twiter y Facebook recopila y analiza datos para el posteriormente ponerlos en funcionamiento con el fin de adaptarse a los clientes, a sus horarios, gustos y necesidades.
- Morgan Stanley: analiza los datos en busca de objetivos financieros y buenas ideas de inversiones para sus clientes.

2.13. CRECIMIENTO

Las barreras al crecimiento de Big Data las impone en gran medida la moralidad de sus usuarios. Se ha descubierto que con las correlaciones de datos se puede predecir en algunos casos con más o menos acierto las acciones futuras. Esto genera cierta incertidumbre en cuanto a legalidad y moralidad como vimos en apartados anteriores.

Respecto a la cantidad de datos que se puedan seguir analizando, se prevén crecimientos importantes, “cloud computing” e “internet of things” avalan que esto seguirá siendo así y el crecimiento de datos no menguara. De forma contraria, se espera que el crecimiento de datos a medida que las nuevas tecnologías avanzan sea cada vez mayor y tenga un crecimiento exponencial.

La implantación de nuevas tecnologías y aplicaciones, requieren de un tráfico, análisis y almacenamiento de datos masivo, a la vez de un ancho de banda cada vez mayor con el que se pueda trabajar con fluidez.

Además la continua implantación de sensores, cámaras, chips y demás mecanismos y el interés del ser humano de conocer y mejorar lo existente es una base fiable para pensar que el manejo de datos crecerá y con ello crecerán la inquietud y nuevas utilidades que se le pueden dar a estos.

Se debe añadir de igual modo que los datos en ciertos casos pierden valor con el tiempo. A pesar de que son muchas las compañías que son reacias a eliminarlos por

completo. Aparte de contar en muchas ocasiones con espacio limitado para el almacenamiento masivo, no siempre son útiles pasado un tiempo desde su captación. Puede imaginarse un valor de la bolsa al que se pretende dar uso para invertir capital, este dato no valdrá de nada pasado un tiempo en el que los valores han cambiado, ya que tienen un tiempo de vida muy limitado. Aunque también es cierto que algunas compañías pueden utilizarlo para generar estadísticas y detectar fluctuaciones en el mercado. Pero para la mayoría será un valor inútil del cual sólo podrán sacar beneficio vendiéndolo a quien esté interesado.

2.14. ERRORES

Como se ha visto anteriormente cabe la posibilidad de que se cometan errores al analizar la información resultante del procesamiento de los datos.

El procesamiento de los “Big Data” se realiza mediante correlaciones, no tiene en cuenta el porqué o el cómo, solo tiene en cuenta el “qué”. Es decir, puede darse una correlación entre datos que aparentemente no tenga sentido alguno y que parezca un error. Pero realmente no tiene por qué ser un error, puede ser algo cierto que guarda un patrón difícil de reconocer pero existente. Un ejemplo de ello puede ser el extraño caso visto anteriormente de las curiosas ventas en Wal-Mart Estados Unidos. Pero en otros muchos casos estas correlaciones solo serán ruido, sin sentido ninguno que forme parte de las desventajas de analizar grandes cantidades de datos o de haber elegido mal los datos a analizar conjuntamente. Para evitar esto se precisa de un personal experto y formado en la materia, que sea capaz de separar la información importante del ruido con destreza.

Aunque en todo calculo estadístico puede darse ruido, soluciones falsas o basura, es más probable que esto ocurra en el procesamiento de Big Data. En ello se aumentan los números de datos a tratar y la cantidad de variables sobre las que se han de calcular correlaciones. Esto produce estadísticas falsas que siguen una forma convexa similar a la imagen. La gráfica es creciente a medida que crece el número de variables a analizar, por lo tanto Big Data requiere de una importante fase de depuración, ya que no queda libre de errores y correlaciones poco realistas aunque existentes.

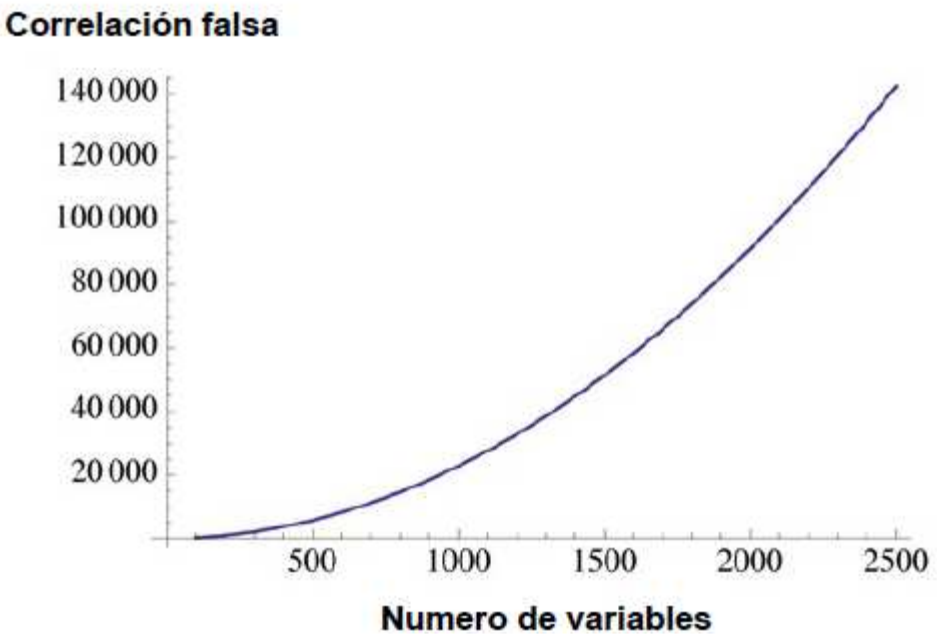


Ilustración 6. Relación entre numero de variables y de correlaciones falsas

3. HADOOP

En este capítulo se estudia el funcionamiento de la plataforma Hadoop y de las diferentes plataformas con las que coopera hasta obtener un sistema completo, con una gran capacidad de resolución a los inconvenientes que puedan encontrarse durante su uso.

Hadoop se define como un framework de licencia libre, escrito en Java y concebido para sistemas basados en Linux. Su objetivo principal es el almacenamiento y procesamiento de grandes cantidades de datos de distintas naturalezas, en un sistema que puede ser distribuido y que trabaja a una baja latencia.

Ilustración 7. Logotipo Hadoop

Antes de profundizar más en Hadoop y en el conjunto de software que le acompaña, se presentan a continuación las definiciones y usos de algunas tecnologías íntimamente ligadas a Hadoop.

3.1. OPEN SOURCE

Se entiende como Open Source o código abierto, aquel software que es libremente distribuido. Es decir, aquel del que se dispone la totalidad de su código y puede ser modificado por otros que no sean el propietario original.

La creación de software Open Source se inicio aproximadamente en el año 1983, mediante el proyecto GNU. Surgió debido a la necesidad del momento, ya que hasta entonces la obtención de software era muy privativa. Esta definición fue reconocida como tal en el año 1998 de la mano de Netscape Communication Corporation.

Open Source ha tenido su mayor índice de crecimiento en los últimos años ya que ha está íntimamente ligado con el mayor avance tecnológico. Con Open Source se ha conseguido no sólo que el software de código abierto llegue a más usuarios, sino que sea de mejor calidad que otros que no son de código abierto. Esto se consigue mediante las aportaciones, grandes o pequeñas, de empresas y programadores.

Actualmente, compañías como HP e IBM, certifican las bases de código de plataformas Open Source, para así darles más fiabilidad y promover mejor su venta y uso.

Concretamente el framework que este texto ocupa, Hadoop, actualmente es coordinado por Apache Software Foundation y es distribuido bajo compañías como Apache, Microsoft y otras, que aportan a esta herramienta valores adicionales. Algunos de estos valores adicionales resultan ser casi imprescindibles para muchos, como por ejemplo la posibilidad de hacerlo funcionar en sistemas operativos Windows y no solo en sistemas basados en Linux como fue en su concepción.

Mediante la libertad de disponer del código fuente del software, diferentes compañías y programadores pueden aplicarle mejoras, ir subsanando defectos o añadiendo nuevas funcionalidades. De este modo un framework como Hadoop se convierte en un autentico ecosistema, capaz de dar solución a una gran variedad de cuestiones y nuevas necesidades que se plantean mientras se usa.

En capítulos posteriores se enumeran algunas de las aportaciones que han ido realizando para conseguir una mayor funcionalidad.

En este momento, un gran número de software que cumple las premisas de Open Source es gestionado y compartido bajo la licencia Creative Commons[9], fundada en el año 2001 y que se encuentra en plena expansión, teniendo cada vez más afiliados a lo ancho de todo el mundo.

3.2. OPEN DATA

Este término viene utilizándose desde aproximadamente el año 2010, y se ha convertido en una línea a seguir por parte de empresas, organismos y gobiernos.

Bajo Open Data se permite la publicación de datos obtenidos por diferentes organismos para ser compartidos de forma pública en la red sin limitaciones. Todos tienen la importante característica, de que al tratarse de datos públicos carecen de identidades personales. Estos datos pueden ser de naturaleza geográfica, médica, atmosférica y medioambiental, matemática, criminal, etc.

Con este movimiento pasan a estar disponibles infinidad de datos para su análisis y procesado. Sectores como Big Data se ven beneficiados muy directamente. Países como España, se han unido mediante la creación de sitios web donde albergar los datos y ponerlos a disposición del público. Concretamente en el caso de España los datos compartidos se pueden encontrar en Datos.gov[10], Alemania también pone a disposición del mundo datos a través de GovData, Estados Unidos mediante la creación del National Intelligence Open Source Center (OSC) y Australia mediante el National Open Source Intelligence Centre (NOSIC).

[9] Más información sobre Creative Commons en el glosario.
[10] Puede obtenerse una lista más detallada de distintas fuentes de Open Data en el glosario

3.3. CLOUD COMPUTING

El término hace referencia al acceso casi inmediato a software y hardware remoto de forma fácil desde cualquier ubicación. Con ello se consigue poder realizar tareas que serían imposibles con el hardware o software disponible físicamente por el usuario.

La existencia de Cloud Computing hace favorable el crecimiento y desarrollo de entornos como Hadoop. Son muchas las empresas y organismos que ante la imposibilidad, por motivos económicos o de espacio, se deciden a "alquilar" el equipo en la "nube" para realizar el almacenamiento y procesamiento de sus datos y pagar únicamente por lo que requieren en cada momento. Mediante estas prácticas no necesitan generar una gran inversión en equipos o en salas de almacenamiento, además de ahorrarse la contratación de personal cualificado que mantenga los equipos en funcionamiento. Esta labor se la dejan a compañías como Google o Amazon que disponen de grandes conjuntos de equipos de gran calidad y del personal cualificado necesario para que su funcionamiento sea óptimo.

Tres de las compañías que permiten alquilar sus servicios y máquinas para hacer funcionar proyectos de Big Data son:

- Amazon Web Services (AWS): plataforma de servicios web ofrecida por Amazon, fue lanzada en el año 2006. Actualmente ofrece diferentes servicios web sobre Hadoop dependiendo de las necesidades del cliente. Con este servicio permite abaratar costes a las empresas en el mantenimiento y adquisición de equipos, a la vez que permite obtener siempre la tecnología más avanzada y un funcionamiento eficiente supervisado por el equipo de Amazon.
  El servicio tiene una facturación dependiente del uso. Este ofrece bajas latencias, seguridad en el tratamiento de los datos, gran velocidad de procesamiento, un equipo software de calidad adaptado al servicio y un centro de soporte.

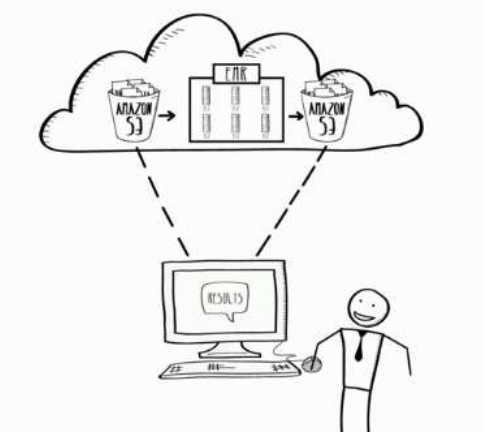


Ilustración 8. Funcionamiento de Amazon Web Services[11].

[11] Imagen tomada de Amazon.com

- Google Cloud Platform: plataforma de servicios web ofrecida por Google, lanzada en 2008 continúa aún en construcción, añadiéndosele cada año nuevos servicios. Para el manejo de grandes cantidades de datos ofrece el servicio BigQuery, con el que analizar los datos en la nube. Además ofrece diferentes interfaces de uso dependiendo del usuario, ya sean administradores, desarrolladores y demás. También ofrece un servicio de asistencia y aprendizaje en red. La facturación del servicio depende del uso.

- Microsoft Azure: plataforma de servicios en la nube ofrecida por Microsoft, anteriormente conocida también por Windows Azure o Azure Services Platform, fue lanzada para su comercialización en el año 2010.

3.4. INTERNET OF THINGS (IoT)

Literalmente traducido por el internet de las cosas. Su idea es conseguir que cualquier dispositivo sea medible mediante la colocación de sensores, chips y demás dispositivos de identificación y medición, para así crear con ellos una subred que permite estar interconectados con los aparatos.

Muchas compañías y cada vez más, se unen a esta tendencia creciente por todo el mundo, dotando a sus artículos de estos sensores, de modo que se puede obtener en cualquier momento su estado.

Internet of Things favorece en gran medida la necesidad de Big Data, ya que se hace necesario almacenar los datos obtenidos de los dispositivos, para luego ser procesados en busca de fallos, posibles mejoras, estadísticas de funcionamiento y demás.

Además, el hecho de que las compañías estén fomentando su crecimiento, no sólo les favorece a ellos mismos, también para el usuario normal. Sin apenas darse cuenta se está convirtiendo en algo cada vez más necesario y útil. Permite a los más adentrados en la tecnología formar una red de comunicación con sus aparatos más utilizados, pudiendo así tener acceso a ellos en cualquier momento. Pero los que menos idea puedan tener sobre esta tecnología también probablemente estén de un modo u otro ligados al internet de las cosas, y es que el más sencillo aparato que se posea en casa puede formar parte del internet de las cosas.

Donde realmente se hace importante la relación del internet de las cosas y el Big Data sigue siendo en las empresas. Muchas de ellas, no utilizan Big Data por la cantidad de datos que registran introducidos manualmente por las personas. Sino que utilizan Big Data por la gran cantidad de datos que obtienen mediante conexiones máquina a máquina, mediante los datos recibidos procedentes de sensores en los aparatos que

fabrican, o mediante con la captación de datos con los que posteriormente realizar estudios para analizar su productividad.

También se relaciona íntimamente el uso de sensores en aparatos convencionales con uno de los objetivos de Big Data, ofrecer a cada usuario lo más "conveniente" en cada momento. El hecho de tener un dispositivo en el tanque de combustible del coche puede hacer que gasolineras cercanas traten de comunicarse con a través del Smartphone o a través del navegador del automóvil para ofrecer sus servicios de repostaje. Incluso el uso de la cafetera por la mañana puede indicar un horario y una disponibilidad a partir de la cual puede llegar publicidad al ordenador o teléfono móvil del usuario, orientada a los gustos del consumidor mientras toma café.

En la actualidad una de las mayores fuentes de información que posee Big Data es la información procedente de las redes sociales. Pero se prevé que esto cambie y que para el año 2015 la mayoría de información utilizable en Big Data provenga del Internet de las cosas.

3.5. ARQUITECTURA GENERAL HADOOP

Hadoop está formado por dos componentes principales. Un modelo de programación conocido como MapReduce o YARN (MapReduce 2.0), que es la capa de procesamiento de datos y HDFS (Hadoop Distributed File Sistem), que es básicamente la capa de almacenamiento de los datos de forma distribuida.

El conjunto de la arquitectura de Hadoop está organizada en nodos distribuidos en diferentes racks.

En los siguientes puntos se estudiará más a fondo cada capa de la arquitectura.

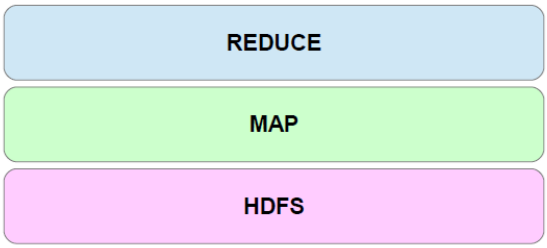


Ilustración 9. Arquitectura Hadoop

3.5.1. MAPREDUCE

Es un paradigma de programación dividido en dos fases: Map y Reduce. Fue creado en el año 2004 por Jeffrey Dean y Sanjay Ghemawat en Google y utilizado por primera vez para calcular el algoritmo de PageRank[12] de Google.

MapReduce hace uso del concepto "divide y vencerás". Tiene la capacidad de dividir una petición por parte de un cliente en otros muchas partes y encargar el trabajo a múltiples nodos que funcionan en paralelo. Actualmente tiene la capacidad de procesar 20PB diariamente con un conjunto desde 2 a 10.000 máquinas.

Esta capa de la arquitectura de Hadoop es el principal encargado de la gestión de recursos y procesamiento de datos, su arquitectura general es de la forma:

- Map: se trata de la función map(), consiste básicamente en el mapeo de la información entrante. Esta fase tiene como entrada la información y como salida un par [clave: valor] que será la entrada de la siguiente fase. Dependiendo de la cantidad de información que tiene a la entrada es capaz de generar varias tareas en paralelo, de modo que logra minimizar la latencia.

(clave1, valor1) →[(clave2, valor2)]

- Reduce: esta fase se trata de la función reduce(). Es la encargada de realizar el procesamiento de la información recibida, ya mapeada en el paso anterior. Tiene como entrada el par [clave: valor] obtenido de la fase anterior y como salida otro par [clave: valor].

(clave3, [valor2]) →(clave2, valor2)

Para el programador o usuario este proceso de Map Reduce es invocado mediante el método JobClient.runJob(conf) perteneciente a la clase JobClient. Éste puede estar implementado en lenguaje Java, Phyton, C++, C y Perl entre otros. Concretamente la versión de Hadoop esta implementada en Java.

3.5.1.1. ARQUITECTURA

Dentro de la arquitectura de MapReduce formada dos bloques Map y Reduce, se encuentran otros elementos que son los encargados de ejecutar las diferentes tareas. Los elementos principales que lo componen son:

[12]PageRank: marca registrada por Google en 1999. Cuyo algoritmo realiza el cálculo de datos para asignar una relevancia determinada a cada página web. Actualmente este algoritmo sigue siendo utilizado por el motor de Google

- Cliente: es siempre el iniciador de la funcionalidad Map/Reduce, ya que es el único elemento existente que tiene la capacidad de poner en funcionamiento el proceso enviando un trabajo.

- Organizador de intercambios en el sistema distribuido: debido a la importancia de distribuir el trabajo, este es un elemento principal. Es el encargado de la distribución de trabajo entre las distintas entidades.

- JobTracker: es el coordinador de todo el trabajo. Su implementación es en Java y su clase principal es JobTracker. Existe un único JobTracker por clúster encargado de recibir todas las peticiones de los clientes y organizar el trabajo para los TaskTrackers. Para la elección del TaskTracker, el JobTracker tiene en cuenta la carga que tengan, es decir su estado, si tienen slot disponibles o no y si estos TaskTrackers se encuentran o no en el mismo rack. Además el JobTracker durante todo el tiempo que los TaskTracker están realizando su trabajo no pierde contacto con ellos. Todos los TaskTracker deben enviar un paquete de control cada varios minutos para tener informado al JobTracker. Éste además posee en todo momento la identificación de cada uno de ellos.

- TaskTracker: son los elementos encargados de realizar las tareas en las que se ha dividido el trabajo mediante la creación de distintos hijos que trabajen en paralelo. Estos pueden estar en tres estados no simultáneos, en reposo, trabajando o terminado. Su implementación en el caso de Hadoop también es en Java y su clase principal es TaskTracker.

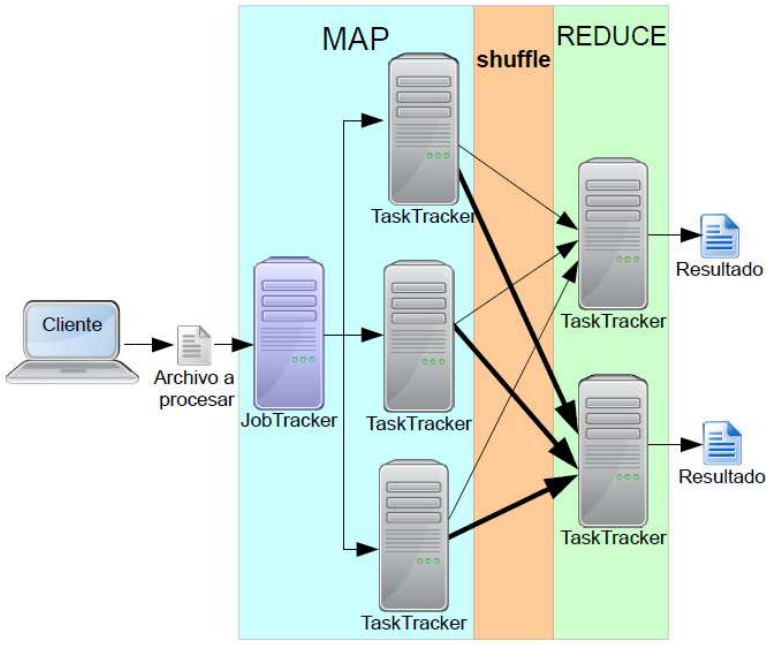


Ilustración 10. Arquitectura MapReduce

3.5.1.2. FUNCIONAMIENTO

Desde que el cliente realiza la petición hasta que esta es completada, los sucesos que se producen son los siguientes:

- El cliente realiza una petición, que es recibida por el nodo JobTracker.

- El nodo JobTracker una vez ha recibido la petición, dependiendo del tamaño de esta calcula si se la debe de asignar a uno o a varios nodos TaskTrackers. Para ello divide la petición en diferentes partes de entre 16 y 128 megabytes (este valor es ajustable por el usuario dependiendo de su preferencia). Tras esto busca tantos TaskTrackers en reposo como necesite, tanto para ejecutar las operaciones map como para las operaciones reduce. Finalmente envía las porciones de trabajo a los elegidos como nodos para la función map.

- Estos TaskTrackers realizan su tarea en paralelo. Van almacenando sus resultados en una memoria cache.

- Tras esto comienza la etapa "shuffle". Todos los valores intermedios obtenidos de los procesos map se organizan teniendo en cuenta la clave que se le asigno en la etapa map y se les envía a los nodos encargados del proceso reduce.

- Estas respuestas se combinan mediante el método reduce

- El resultado final se le devuelve al JobTracker que inicio la comunicación una vez todos los procesos map y reduce han terminado.

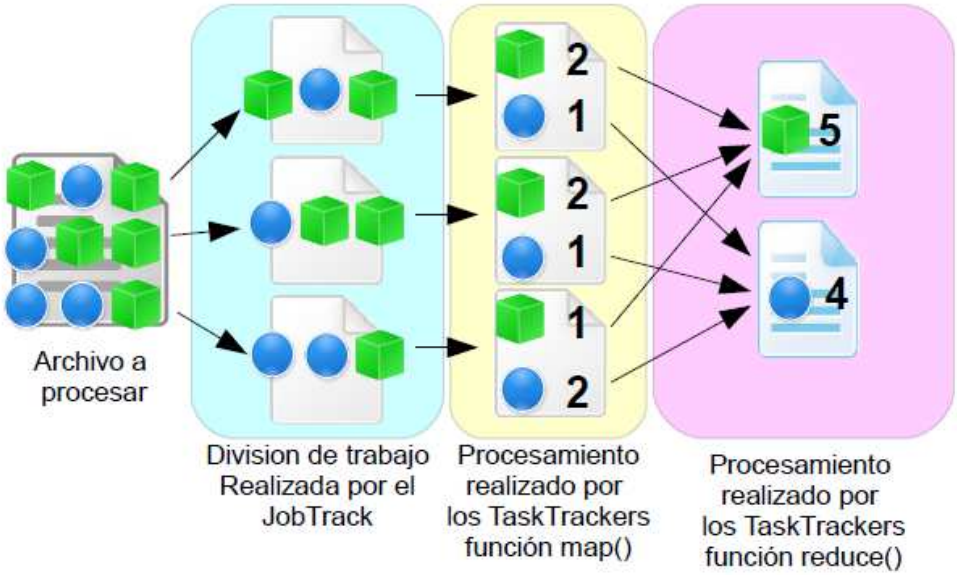


Ilustración 11. Procesamiento archivos con MapReduce

3.5.1.3. PROBLEMAS

Los principales problemas que pueden existir en la funcionalidad MapReduce son la falta de memoria, la caída de un nodo o un hardware con características insuficientes.

Ante los problemas de insuficiencia de memoria o hardware con capacidades de procesamiento insuficientes, MapReduce tiene la capacidad de ser escalable, lo que convierte un problema en una gran cualidad de MapReduce.

Frente a problemas de caídas de nodos TaskTracker, tiene la capacidad de activar otro nodo, que se encuentre en ese momento en reposo, para continuar la operación que el anterior no ha podido ejecutar. Es decir, envía a un nuevo nodo el trabajo que el anterior no ha podido terminar antes de que se cayera, reinicia el nodo caído y le devuelve al estado de reposo, de este modo queda ocioso para nuevas peticiones. Esto es posible debido a que el nodo JobTracker está siempre en continuo contacto con el resto de nodos para así detectar las caídas o fallos que se produzcan.

Un problema más específico es que se produzca la caída del nodo JobTracker, en ese caso ocurre que todo el proceso MapReduce debe de repetirse tras su reinicio y correcto funcionamiento.

Por tanto se puede afirmar que posee una gran capacidad de detección y recuperación ante errores, lo que suma una gran ventaja al sistema Hadoop.

3.5.2. YARN

Es la evolución de MapReduce, conocido como MapReduce 2.0 o YARN (Yet Another Resource Manager) y distribuido por Apache Software Foundation desde el año 2012. Esta evolución representa la mayor diferencia entre lo que se conoce como Hadoop 1.0 y Hadoop 2.0.

El principal cambio respecto a la versión anterior es la separación de las principales tareas que antes realizaba el JobTracker, siendo ahora un modulo separado, el Resource Manager (RM) el encargado de la supervisión y negociación de los recursos. Es básicamente el encargado de organizar el trabajo, tarea que antes realizaba el JobTracker además de encargarse de recibir las peticiones de los clientes.

3.5.3. HDFS

Hadoop Distributed File Sistem, es un sistema de ficheros basado en la distribución de la información en distintas máquinas que pueden ser geográficamente muy distantes, conectadas entre sí mediante una red de modo transparente al usuario.

Al trabajar con una gran cantidad de datos se hace muy importante la velocidad de lectura, ésta al realizarse de forma paralela en distintas máquinas agiliza el proceso, comparado con leer todos los datos de un mismo disco. Además HDFS tiene la ventaja de funcionar en un sistema distribuido, permite leer datos en paralelo de varios equipos, independientemente de su localización geográfica y todas las máquinas son vistas por el sistema como una única máquina de gran capacidad.

Este sistema de ficheros está basado en el sistema GFS (Google File Sistem) creado en 2003, mejora la capacidad de distribución, escalabilidad, portabilidad y tolerancia a fallos, además tiene la capacidad de crecer con la demanda requerida sin mayor impedimento.

Sus principales diferencias con otros sistemas de ficheros distribuidos son que se trata de un sistema independiente de cualquier software o hardware especifico para trabajar. Está diseñado para operar sobre software de bajo costo y otorga cierta permisividad en la normativa POSIX, esto es que sacrifica en cierto modo las normativas para favorecer un mejor rendimiento y manejo. Además le hace un sistema resistencia a fallos, cualidad importante debido al gran número de máquinas con las que trabaja y la suma de probabilidades de que ocurra un fallo en cada una de ellas.

En HDFS los ficheros se almacenan en bloques que no tienen porque estar en el mismo nodo. Tiene la capacidad de equilibrar los bloques en distintos nodos a demanda, por lo que se puede distribuir un fichero en bloques almacenados en distintos nodos sin problema. Tiene la particularidad de que en HDFS los bloques son de un tamaño de entre 64MB y 128MB, superiores a los habituales para minimizar la latencia en los accesos de lectura. Por ello HDFS está indicado para accesos largos y no pequeños accesos que producen desventajas en la latencia.

El fichero entonces queda dividido en bloques del mismo tamaño, con la particularidad de que a diferencia de la forma general, el último bloque no tiene por qué ser de un tamaño fijo y es posible completarlo con relleno. HDFS permite que el último bloque sea del tamaño requerido por el fichero sin límites que alcanzar, y por tanto con un importante ahorro de espacio.

Para evitar pérdidas de información por caídas de nodos, el sistema réplica la información en tres nodos, dos copias en el mismo rack y otra copia en un rack distinto por si lo que se produjera fuera una pérdida de un rack completo. El número de réplicas no es un parámetro fijo, puede ser configurado por el administrador.

En el sistema tiene un papel fundamental la seguridad, la ubicación de los nodos y bloques, la tolerancia a fallos y a perdidas de información, por ello trabaja con redundancia y posee una gran tolerancia a nodos caídos en el sistema. Además permite la recuperación del sistema a una situación anterior estable en caso de fallo,

mediante la creación de instantáneas donde almacena el estado del sistema periódicamente o bajo la petición del administrador.

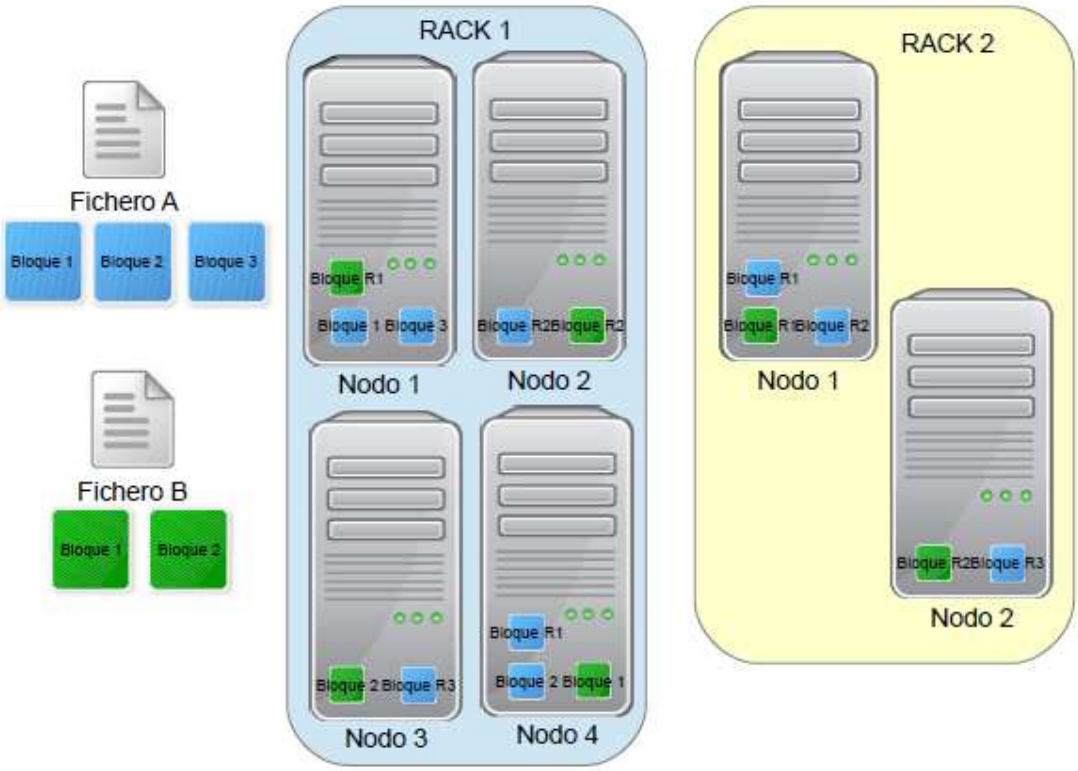


Ilustración 12. Conversión de ficheros a bloques y replicado en diferentes Racks

3.5.3.1. ARQUITECTURA

Un clúster HDFS está formado por dos tipos de nodos:

- Datanode (slave): contienen los bloques de información. Tienen capacidad para la realización de varias tareas simultáneas. Son los encargados de los procesos de lectura/escritura. En algunos casos estos DataNodes son réplicas de otros. En un sistema de archivos vienen representados por dos ficheros, uno que contiene los datos y otro que contiene los metadatos y el checksum. Todos los DataNodes siempre están identificados por un ID de almacenamiento que les caracteriza frente a un NameNode para permitir su registro y acceso. El sistema soporta entre 10 y 4000 DataNodes.

- Namenode (master): es el nodo encargado del cierre, apertura y renombrado de directorios y ficheros. Además contiene los espacios de nombres del fichero y por lo tanto un árbol con la topología que siguen el resto de nodos. Es también el encargado de la gestión de los metadatos y de la gestión y coordinación de los bloques de datos. Posee el control y la información sobre la asignación de los bloques en los DataNodes, así como de la creación, eliminación, movimiento, renombrado de directorios y supervisión del número de réplicas existentes y de su estado. En caso de pérdida de

una réplica será el encargado de replicarla de nuevo en otro DataNode. En el Namenode los ficheros y directorios están representados por inodos que contienen información sobre permisos, modificaciones, últimos accesos, espacio de nombres y cuotas de disco. Tiene la capacidad de generar una "imagen" que contiene los inodos y la lista de bloques que definen los metadatos del sistema de nombres. Estos mantienen una imagen de todo el espacio de nombres en memoria RAM para facilitar su acceso y soportan la conexión simultánea de miles de DataNodes y conexiones de clientes; aunque estos últimos producen cierto "cuello de botella", por lo que el Namenode realiza las transacciones de los clientes por lotes. De este modo cuando varios hilos tienen que iniciar una sincronización o salvado se agrupan en un lote, de modo que un solo hilo se encarga de toda la transacción y el resto de hilos esperan para comprobar que sus operaciones han sido realizadas con éxito. Con este método se ahorran gran cantidad de operaciones de flujo desde el NameNode a los DataNodes. Existe un único nodo de este tipo por clúster.

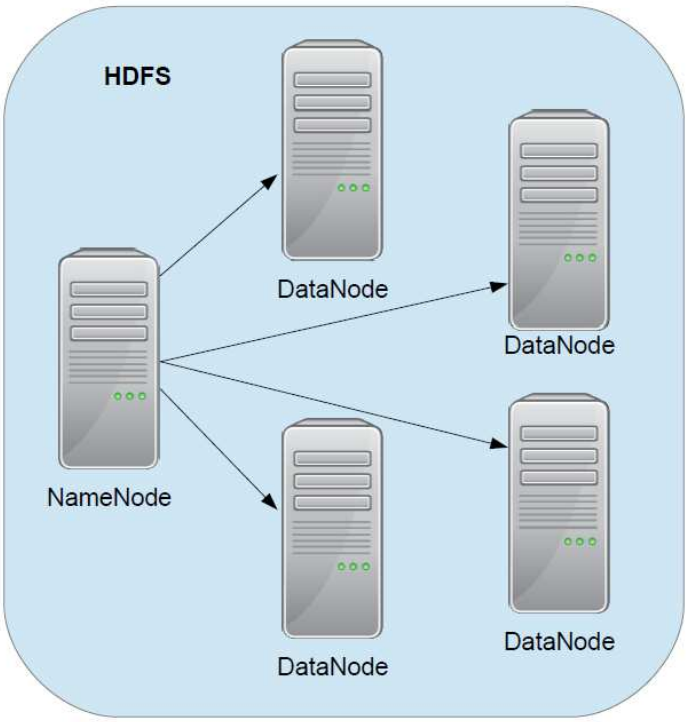


Ilustración 13. Arquitectura HDFS

Otros:

- Cliente: accede al sistema mediante la interfaz de HDFS únicamente. Posee la capacidad de creación de aplicaciones, como se verá más adelante, en diversos lenguajes de programación fácilmente adaptables a las características de Hadoop, mediante las cuales acceder y manejar los datos. El cliente además dispone de las operaciones de lectura, escritura y eliminación de ficheros y creación y eliminación de directorios, siendo para el usuario totalmente invisibles las acciones que tiene el NameNode con los DataNodes.
  Todas las transacciones iniciadas por un cliente hacia el NameNode se almacenan en un "journal", de modo que se tiene un registro de todas las operaciones.
- Checkpoint Node: almacena la "imagen" generada por el NameNode que contiene los inodos y la lista de bloques que definen los metadatos del sistema de nombres.
- Backup Node: es el nodo encargado de la creación de Checkpoints periódicos con el espacio de nombres, mantiene esta información en memoria por si el NameNode fallara, tener el último estado en el que se encontraba. Básicamente tiene todas las mismas funciones que el NameNode con la diferencia de que no puede modificar el espacio de nombres ni conocer la ubicación de los bloques.

Los ficheros normalmente utilizados son los denominados "Write once read many", es decir, los ficheros se escriben una vez y son leídos en múltiples ocasiones. Este es el funcionamiento común, ya que se suele tratar de la grabación de una gran cantidad de datos y su lectura en varias ocasiones por parte de los usuarios con el fin de realizar distintos estudios.

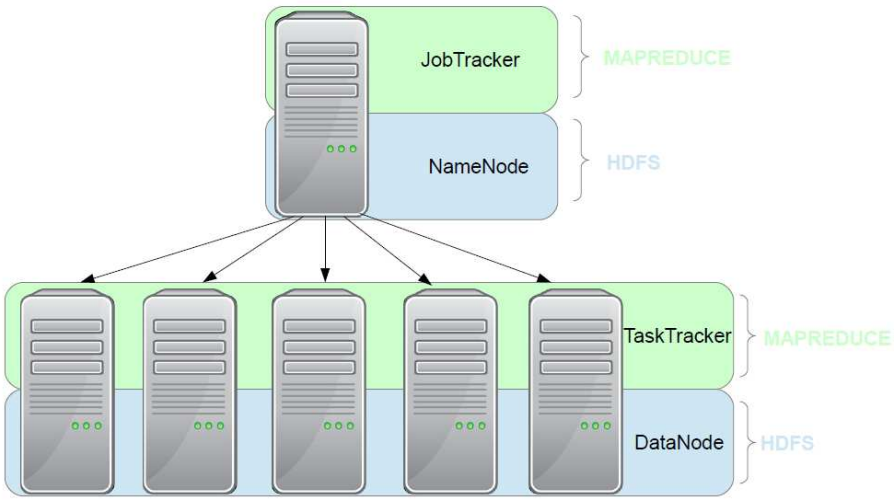


Ilustración 14. Arquitectura HDFS Y MAPREDUCE

3.5.3.2. CARACTERÍSTICAS

Cuando el sistema arranca, cada DataNode realiza un intercambio de datos con el NameNode para comprobar el identificador de cada uno en el espacio de nombres y si las versiones de software son idénticas. En caso negativo, en el que no coincida cualquiera de los datos, el DataNode es apagado automáticamente. Al impedir la conexión de un DataNode con el NameNode sin coincidir el identificador se protege la integridad del sistema de archivos.

En el caso de que se desee integrar un nuevo DataNode al sistema por primera vez y este carezca de identificador registrado en el NameNode, se requiere que el NameNode reciba un identificador y posteriormente vuelva a realizar la conexión con el NameNode y se proceda al registro.

El identificador de almacenamiento de cada DataNode permite que este sea reconocible por el NameNode, incluso si se han producido cambios geográficos o la conexión se realiza desde otra IP o puerto diferente al habitual. Este identificador es asignado cuando se registra en el NameNode por primera vez y nunca cambia.

Los DataNodes realizan envíos frecuentes de informes de bloques al NameNode para identificar los bloques replicados. En estos informes se envían el identificador del bloque, un sello de creación y la longitud de cada bloque replicado. Este informe es enviado por primera vez cuando se realiza el registro del DataNode y continúan enviándose cada hora, permitiendo así que el NameNode tenga siempre una versión actualizada de las réplicas existentes en el clúster.

Además los Datanodes envían cada tres segundos al NameNode información en la que se incluye el número de transferencias que tienen en curso, la capacidad de almacenamiento y el almacenamiento usado. Mediante esta acción el NameNode tiene siempre una información reciente sobre su actividad y de este modo se le permite generar estadísticas sobre equilibrio de carga y asignación de bloques, a la vez que se mantiene un contacto continuo que indique su disponibilidad. El NameNode en caso de no recibir esta información cada tres segundos del DataNode espera hasta alcanzar los diez minutos, después de no haber recibido de nuevo ninguna información procedente del DataNode, pasa a identificarlo como fuera de servicio y realiza nuevas réplicas de los bloques que contenía.

Los envíos que realizan los DataNodes al NameNode periódicamente son aprovechados por los Namenodes para realizar el envío de instrucciones sobre eliminación o nuevas réplicas, nuevos registros, solicitar informes u ordenar apagado. Con estas instrucciones se mantiene la integridad del sistema y no se compromete su funcionamiento, realizando la comunicación de este modo, el NameNode nunca envía solicitudes a los DataNodes, solo aprovecha sus envíos periódicos para comunicarse con ellos.

Como se mencionó anteriormente el diseño de este sistema esta creado para que los ficheros sean escritos de una vez en un fichero y consultados en múltiples ocasiones. No está optimizado para la modificación frecuente de los archivos, sólo para la consulta, ya que su funcionamiento clásico es la creación de un archivo por parte del cliente, la escritura en el mismo y la lectura en múltiples ocasiones incluso por distintos clientes. Puede darse el caso en el que se requiera por parte del cliente escribir en un archivo creado anteriormente, es decir la modificación del mismo, cuyo proceso se estudiará en el siguiente punto.

3.5.3.3. FUNCIONAMIENTO

El funcionamiento clásico es la creación y escritura de un archivo por parte de un cliente y su lectura en múltiples ocasiones. También puede darse la necesidad por parte de un cliente de la modificación y cierre de un clúster.

A continuación se muestran las fases de cada una de las operaciones.

El modo en el que se atiende una petición de un cliente de creación y escritura pasa por las siguientes fases:

- La aplicación cliente solicita al NameNode la lista de DataNodes disponibles.
- El NameNode facilita la información a la aplicación cliente, al estar el fichero compuesto de bloques, el NameNode cuando lo crea le asigna un bloque con un identificador único y una lista de DataNodes para el propio archivo y para sus réplicas.
- La aplicación cliente contacta con el DataNode correspondiente solicitando el bloque deseado.
- El DataNode devuelve el bloque solicitado para que sea posible la interactuación con el mismo.
- Si la aplicación cliente desea escribir en el bloque, debe de crear una tubería por la que enviar los datos al DataNode. En caso de que la operación llene el bloque, el DataNode debe solicitar nuevos DataNodes, que no tienen porque ser contiguos, para terminar su escritura. Mientras se está realizando la escritura, los Datanodes, donde se almacena el archivo y las réplicas del mismo, crean entre sí una tubería por la que transportar los paquetes mediante el método de ventana deslizante.
- La aplicación cliente al finalizar la escritura y antes de cerrar el archivo genera un checksum por cada bloque, lo que sirve para en el futuro comprobar su integridad.

Ilustración 15. Operación escritura

El modo en que se atiende una petición de un cliente para modificar un archivo pasa por las siguientes fases:

- La aplicación cliente solicita al NameNode la información sobre el DataNode y el bloque correspondiente al archivo que desea modificar.
- El NameNode facilita la información solicitada por el cliente con la particularidad de que permite la modificación del archivo por parte del cliente de forma exclusiva. Mientras que el cliente se encuentra escribiendo en el mismo, se produce un bloqueo y ningún otro cliente puede acceder a escribir en él, aunque si continúan teniendo acceso a su lectura.
- La aplicación cliente escribe en el archivo, y tiene el bloqueo que le permite la escritura mientras haga un envió periódico de informes al NameNode que lo mantenga bloqueado. En caso de que el cliente sobrepase el tiempo asignado de acceso exclusivo, sin haber enviado un nuevo informe y sin haber concluido y cerrado el archivo, otro cliente puede solicitar el acceso de escritura al mismo archivo y quedar bloqueado para el primer escritor. También aunque no haya sido solicitado por ningún otro cliente, el NameNode puede finalizar la exclusividad de escritura al encontrar al cliente inactivo por falta de envíos de

informes periódicos renovando así su exclusividad. De este modo se protege al sistema de posibles bloqueos.

- Otra aplicación cliente puede solicitar el acceso al mismo archivo para una operación de lectura, a lo que el NameNode le permite el acceso, pero a una réplica que no esté siendo modificada en el momento por otro cliente.
- El NameNode, dependiendo de los hechos anteriores, finaliza o continúa permitiendo el acceso exclusivo al archivo antes de cerrarlo.

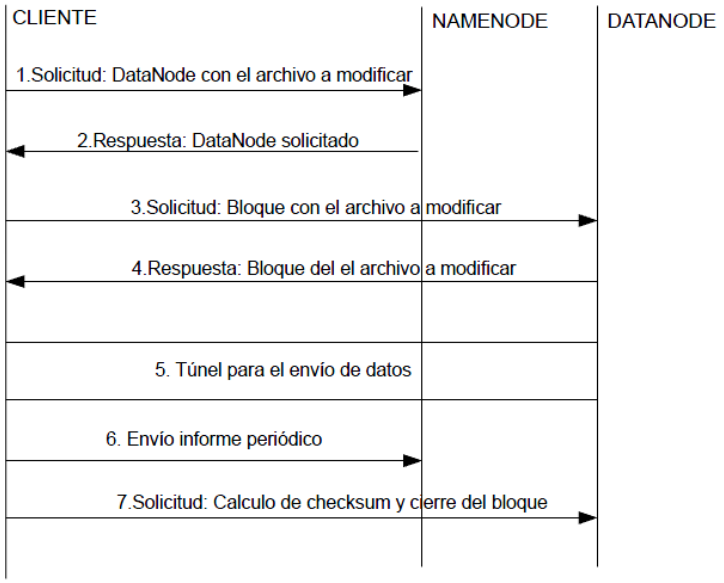


Ilustración 16. Operación modificación. Caso 1: un cliente modifica un archivo

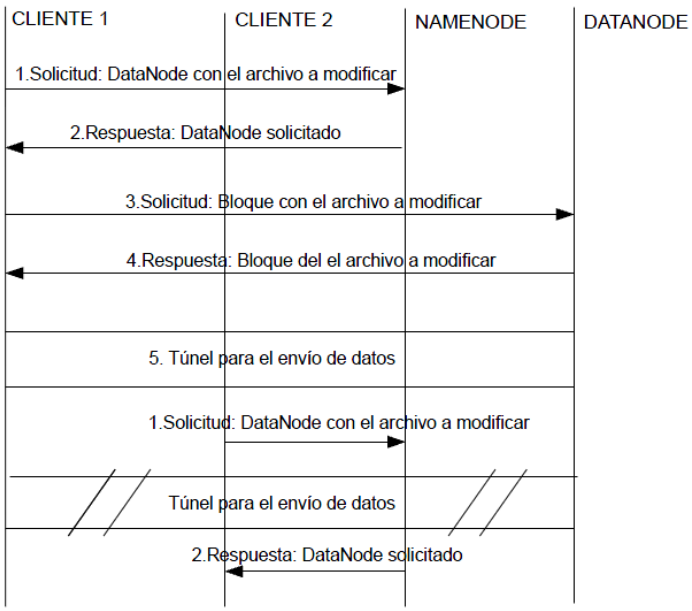


Ilustración 17. Operación modificación. Caso 2. Un cliente modifica archivo sin enviar informes periódicos mientras otro cliente solicita la modificación del mismo archivo

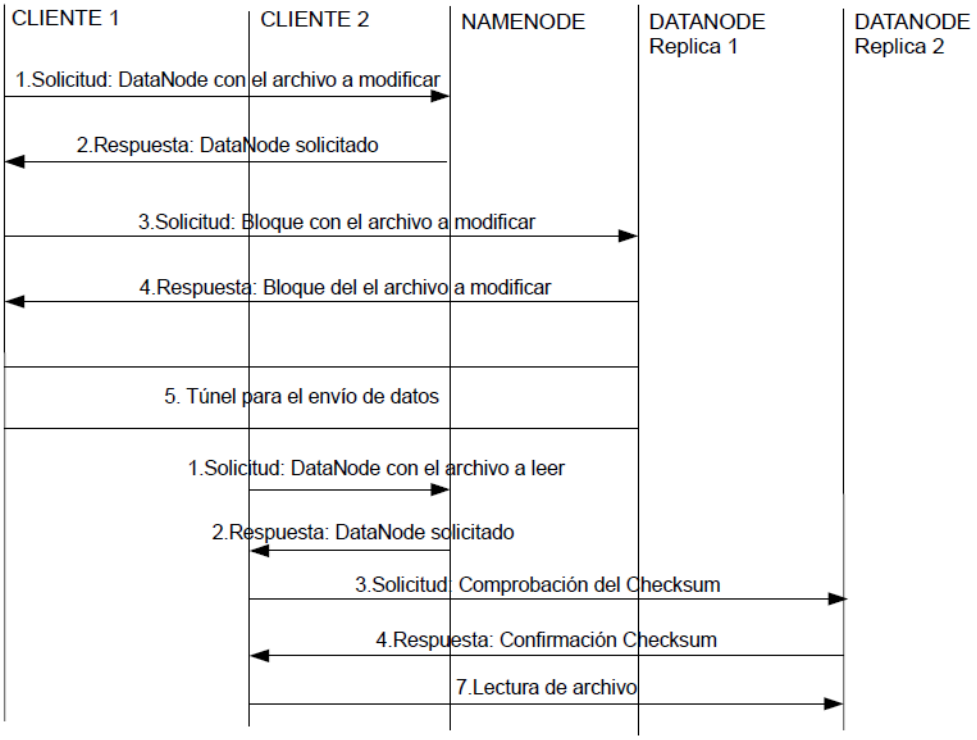


Ilustración 18. Operación modificación. Caso 3: modificación de un archivo por parte de un cliente mientras otro realiza la lectura del mismo archivo procedente de una réplica sin modificar

El modo en que se atiende la petición de un cliente para la lectura de un archivo es:

- El cliente solicita al NameNode un archivo para realizar su lectura.
- El NameNode proporciona al cliente el bloque dependiendo de su ubicación, ya que al disponer de tres réplicas en distintas ubicaciones le ofrece la más cercana.
- La aplicación cliente antes de iniciar la lectura comprueba el checksum del archivo para verificar su integridad. En caso de encontrar alguna disparidad se lo hace saber al NameNode, que es el encargado de catalogar a la réplica como corrupta, iniciar la creación de una nueva réplica y facilitarle la ubicación de la siguiente réplica disponible.

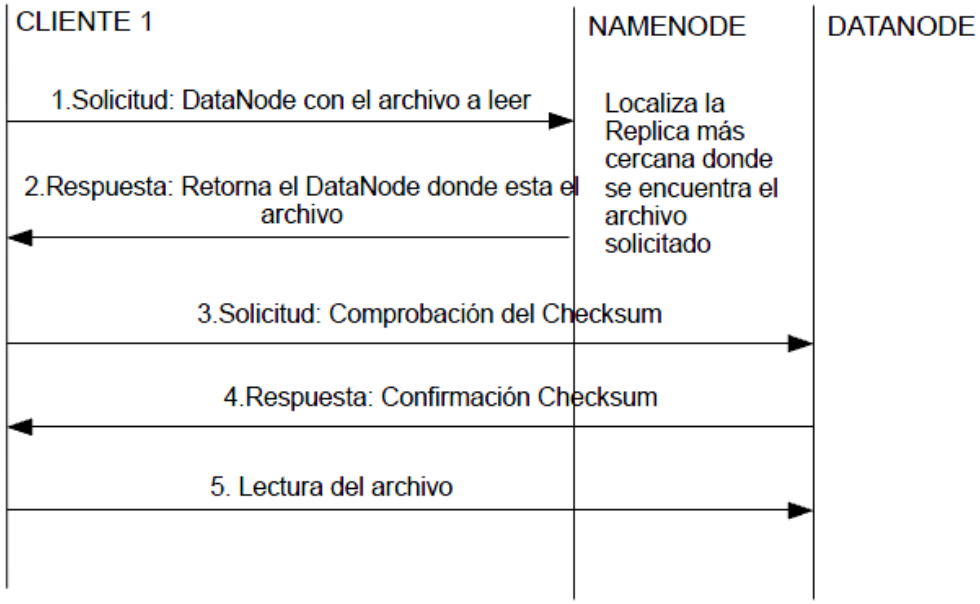


Ilustración 19. Operación lectura. Caso 1: Cliente solicita la lectura de un archivo

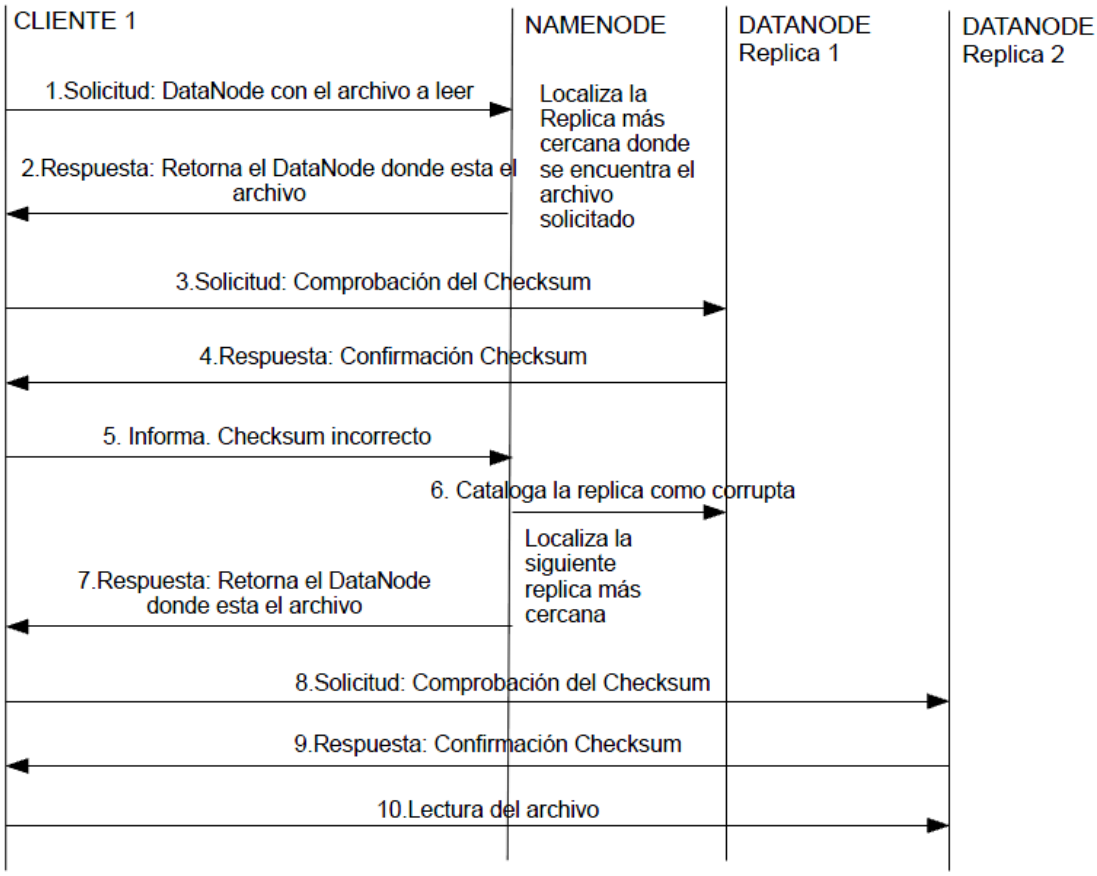


Ilustración 20. Operación Lectura. Caso 2: Cliente solicita la lectura de un archivo que esta corrupto

El modo en el que se atiende la petición de cierre de clúster por la aplicación cliente:

- El cliente solicita al NameNode la petición de cierre total del clúster.
- El NameNode al recibir la solicitud comienza a replicar todos los DataNodes que tenga el clúster albergados a otros clústeres disponibles.
- Una vez el NameNode finaliza la replicación de todos los DataNodes, cierra el clúster finalizando así la operación solicitada por el cliente.

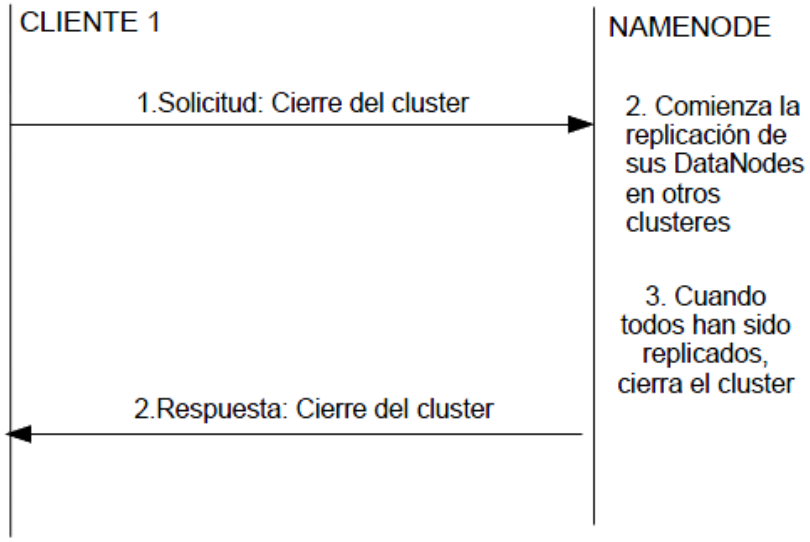


Ilustración 21. Operación: Solicitud por parte de un cliente para el cierre del clúster

3.5.3.4. ERRORES

En el sistema pueden darse errores en el momento de acceder a un DataNode. Estos fallos pueden venir dados porque un bloque ya no se encuentre en el DataNode buscado o que el nodo ya no tenga almacenado el DataNode que contenía el bloque, o que los existentes sean corruptos, etc. Para ello es de vital importancia el checksum de cada bloque, ya que permite que el propio cliente detecte problemas fácilmente y que éstos puedan ser comunicados al NameNode, que es el encargado de realizar nuevas réplicas si lo considera necesario y de realizar los cambios convenientes en caso de caída de un nodo o un bloque.

Para minimizar en la medida de lo posible que esto llegue a ocurrir, se procura siempre facilitar a la aplicación cliente un bloque del que se tiene una constancia reciente de su integridad. Para este fin los DataNodes son escaneados periódicamente, se verifican sus checksums y éstos son almacenados en un informe para tener siempre constancia del último momento en que se verifico cada DataNode y por lo tanto saber cuáles son los últimos que han sido revisados.

3.6. CONFIGURACIONES

Hadoop permite tres tipos de configuraciones:

- Pseudo-distribuida: configuración que permite la simulación de un clúster con varios nodos. Permite simular los diferentes nodos procesando cada demonio en un proceso java diferente.

- Standalone (single-node): es una configuración básicamente para desarrollo y pruebas. Permite la instalación y funcionamiento de un clúster de varios nodos simulados en una máquina virtual. En realidad se trata de un único proceso java para todos los demonios.

- Distribuida (multi-node): configuración típica de un funcionamiento real, permite la configuración de un clúster de varios nodos en distintas máquinas reales.

3.7. COMUNICACIÓN

La comunicación entre los diferentes nodos se realiza mediante el protocolo TCP/IP haciendo uso de la funcionalidad RPC[13].

El cliente es siempre el iniciador de la comunicación, este comunica con el NameNode mediante la previa conexión con un puerto TCP.

La conexión entre nodos requiere un ancho de banda que crezca linealmente con el número de nodos y discos disponibles en el sistema HDFS.

3.8. LIMITACIONES

Tras enumerar las ventajas de Hadoop sobre los sistemas anteriormente conocidos como las bases de datos relacionales, se hace difícil imaginar cuál es la mayor limitación de Hadoop. Posee la capacidad para trabajar sobre hardware económico, es de código abierto, escalable, distribuido y permite el análisis de diferentes naturalezas de datos. Pero no se tiene en cuenta que para muchos objetivos es fundamental el procesamiento en tiempo real, y es esa la mayor limitación de Hadoop.

Hadoop funciona con un procesamiento por lotes, por lo que por sí solo no tiene gran capacidad como herramienta para el procesamiento de datos en tiempo real. Esta limitación se mejora en gran medida cuando Hadoop trabaja apoyada en otras herramientas que si están diseñadas para tal propósito.

[13] (Remote Procedure Call) Protocolo que abstrae al usuario del funcionamiento interno de la comunicación entre máquinas remotas, puede usarse sobre las capas superiores del protocolo TCP

Storm es una herramienta de la compañía Apache, que se explica posteriormente con más detalle. Está diseñada para el procesamiento en tiempo real y permite trabajar conjuntamente con Hadoop, de este modo logran minimizar las limitaciones de Hadoop en gran medida, pudiendo obtener así una plataforma mucho más completa.

Además son muchas las compañías que ofrecen soluciones finales que están tratando de dar una solución a la limitación de Hadoop por medio de otras herramientas. Esto hace que se empiece a vislumbrar el hecho de utilizar Hadoop en un futuro cercano como herramienta para analizar datos en tiempo real, sin verlo como una limitación aún bastante importante.

Otra limitación de Hadoop es su diseño "Write once, read many". Para muchos usuarios de Hadoop es una gran desventaja la dificultad con la que pueden escribir con frecuencia en ficheros dirigidos por Hadoop. El hecho de que esté diseñado para que sus ficheros sean escritos de una sola vez y puedan ser leídos en múltiples ocasiones por distintos usuarios simultáneamente genera un gran inconveniente. Aunque bien es cierto que esta limitación no prohíbe poder modificar un fichero existente. Pero si debería de ser una operación poco utilizada para sacar el verdadero partido a la herramienta y a su capacidad de trabajar por lotes.

Como en el caso anterior, no son tanto unas limitaciones como tal, sino unas restricciones impuestas en el diseño. Hadoop no está diseñado para trabajar con archivos pequeños de datos y accesos frecuentes. Realmente se obtiene todo su beneficio trabajando con grandes cantidades de datos, realizando accesos de lecturas o escrituras largas, donde el tiempo de latencia es mínimo comparado con múltiples accesos breves.

3.9. FUNCIONALIDADES ADICIONALES

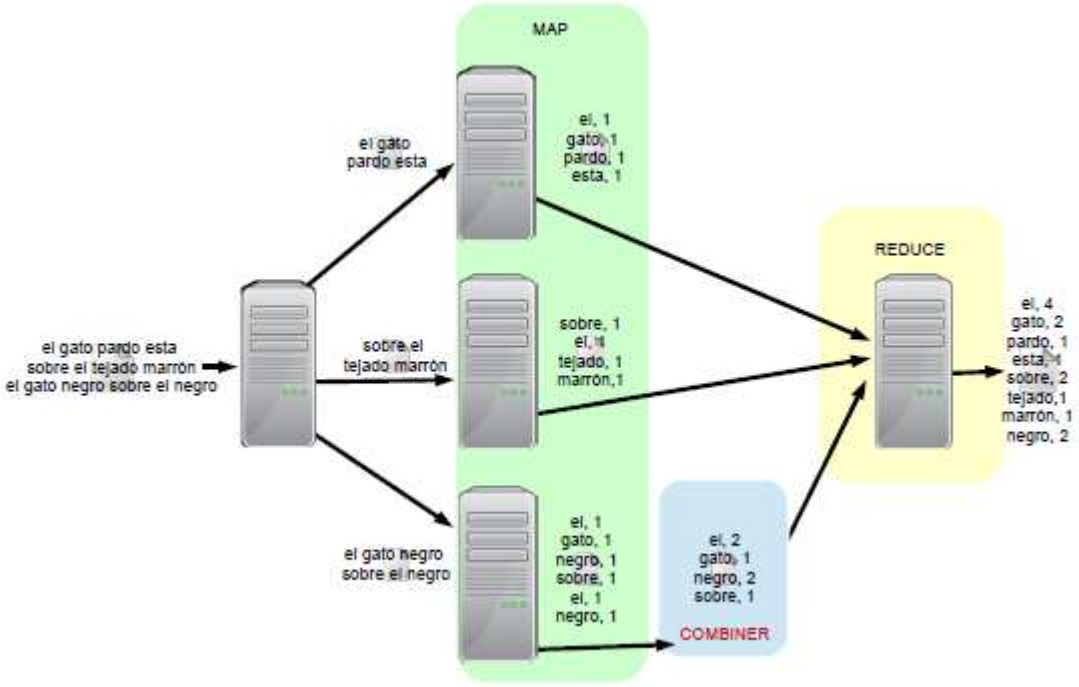
3.9.1. COMBINER

Es una herramienta que permite una mayor optimización en la fase MapReduce mediante el uso de la clase combiner(). Realmente se trata de una función que implementa la interfaz reduce y que es utilizada entre las fases map y reduce.

Su función es optimizar el proceso añadiendo un nueva fase reduce al sistema, de este modo se mejora el procesamiento y permite disminuir la cantidad de datos que tienen que ser procesados en la fase reduce.

Esta función combiner se ejecuta en cada nodo donde anteriormente se ejecutó la función map(). Tras ello envía su salida como entrada a la función reduce(), por lo tanto la salida de las funciones map() no son enviadas a reduce() sino a combiner() y

las salidas de estas si son las redirigidas a la función reduce(). En la imagen siguiente puede verse su funcionamiento:



Con este sencillo proceso se consigue minimizar el ancho de banda y optimizar el tiempo de procesado, aunque no permite ser usado en todos los casos, ya que en muchos la propia función reduce() realiza el mismo trabajo.

3.9.2. CASSANDRA

Base de datos NoSQL[14] de código abierto, desarrollada por Apache Software Foundation con un lenguaje propio CQL (Cassandra Query Language) similar a SQL[15].

Es una base de datos de alto rendimiento. Trabaja de forma que todos los nodos tienen la misma jerarquía, no hace uso de maestros y esclavos, permite la replicación de elementos, trabajar con múltiples máquinas y un fácil escalado.

Ilustración 22. Logotipo Cassandra

[14] Not Only SQL, sistema de gestión de bases de datos por el cual éstas no tienen por qué seguir un esquema entidad-relación.
[15] Structured Query Language, lenguaje de acceso a bases de datos.

3.9.3. HBASE

Base de datos distribuida, no relacional, NoSQL, creada por Apache como código abierto para su uso sobre la capa HDFS.

Esta base de datos tiene la capacidad de trabajar con grandes cantidades de datos almacenadas en un gran número de filas y columnas proporcionando una baja latencia en operaciones de lectura/escritura.

Ilustración 23. Logotipo HBase

3.9.4. HIVE

Software creado por Apache Software Foundation que permite la realización de consultas a los datos de Hadoop mediante su propio lenguaje HQL muy similar a SQL.

Funciona en la capa superior a Hadoop y esta optimizado para realizar lecturas en la base de datos, pero no para realizar una gran cantidad de operaciones de escrituras. Tampoco es recomendable cuando se requiere una latencia baja.

Ilustración 24. Logotipo HIVE

3.9.5. PIG

Herramienta con entorno interactivo y con lenguaje propio de programación denominado Pig Latin creado por Yahoo!.

Su tarea es facilitar el uso de Hadoop a cualquier usuario centrándolo en el procesado y análisis de la información y menos en el desarrollo de programación para manejar Hadoop.

Ilustración 25. Logotipo Pig

3.9.6. AMBARI

Plataforma creada por Apache que permite trabajar con Hadoop mediante una interfaz grafica intuitiva.

Mediante esta plataforma se permite al usuario trabajar haciendo uso de las herramientas Hive, HBase, Pig, Sqoop, Zookeeper y Oozie además de con la arquitectura típica de Hadoop, HDFS y MapReduce.

Mediante Ambari el administrador puede crear clústeres Hadoop, realizar el mantenimiento y administración y finalmente realizar el monitoreo necesario.

3.9.7. AVRO

Herramienta de código abierto creada por Apache Software Foundation.

Se trata de un sistema de serialización de datos basados en esquemas JSON[16], con ello permite una representación de estructuras complejas. Sus datos son almacenados en formato binario, lo que consigue hacerlo una herramienta liviana y eficiente.

Ilustración 26. Logotipo Avro

[16] JavaScript Object Notation, formato para el intercambio de datos. Es una alternativa al formato XML que funciona con todos los lenguajes de programación.

3.9.8. CHUKWA

Subproyecto de Hadoop que trabaja sobre las capas HDFS y MapReduce, por tanto adquiere su robustez y escalabilidad.

Concebida para encargarse de la recopilación y análisis de registros a gran escala.

Ilustración 27. Logotipo Chukwa

3.9.9. MAHOUT

Librería de código abierto creada por Apache Software Foundation.

Esta librería esta implementada en Java como mecanismo de aprendizaje automático para máquinas con interfaz de línea de comandos.

Con esta herramienta se obtienen multitud de algoritmos que implementan casos típicos en una gran variedad de bibliotecas, entre ellas se encuentran bibliotecas específicas para el tratamiento de determinadas topologías de clústeres y nodos, bibliotecas matemáticas y estadísticas.

Ilustración 28. Logotipo Mahout

3.9.10. SPARK

Herramienta de código abierto que complementa a Hadoop, desarrollada por Apache Software Foundation con la colaboración de otras empresas como Yahoo! e Intel.

Trabaja sobre la capa HDFS, permite mayor rendimiento que MapReduce y propicia el aprendizaje automático de algoritmos a la máquina.

Ilustración 29. Logotipo Spark

3.9.11. ZOOKEEPER

Es un servicio centralizado creado por Apache.

Permite mantener la coordinación entre las distintas máquinas distribuidas, mediante una sencilla y configurable interfaz se pueden gestionar jerarquías en los nodos y hacer que trabajen con coherencia obteniendo un mayor rendimiento.

Ilustración 30. Logotipo Zookeeper

3.9.12. HADOOP STREAMING

Utilidad API incluida en la distribución de Hadoop. Permite el procesamiento de varias líneas de texto al contrario que la API básica de Hadoop, ya que ésta controla los mecanismos de lectura. Además permite la escritura de los trabajos MapReduce en diferentes lenguajes, además de en Java.

3.9.13. SQOOP

Herramienta creada por Apache Software Foundation.

Funciona como intermediario entre Hadoop y bases de datos relacionales como Hive o HBase mediante una interfaz de línea de comandos.

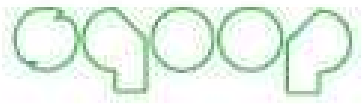


Ilustración 31. Logotipo Sqoop

3.9.14. STORM

Herramienta para el tratamiento y análisis de información en un sistema distribuido en tiempo real diseñada por Apache.

Esta herramienta es de código abierto y permite su uso en múltiples lenguajes de programación, es fácilmente escalable y debido a su capacidad de funcionamiento con colas y bases de datos le permite ser una herramienta auxiliar a Hadoop para trabajar en tiempo real.

3.9.15. OOZIE

Proyecto de código abierto creado por Apache.

Se trata de un planificador de flujos de trabajo para Hadoop. Esta herramienta permite cierta independencia al sistema Hadoop, gestionando los trabajos según avance el progreso del trabajo y la demanda. Permite programar trabajos que se activan dependiendo de la salida del flujo anterior cuando este termine.

Además permite al administrador programar un flujo de trabajos en secuencia dependientes entre sí o no y en cadena o programar un flujo de trabajo en intervalos de tiempo regulares.

Ilustración 32. Logotipo Oozie

3.9.16. FLUME

Herramienta creada por Apache.

Su objetivo de permitir de forma eficiente la recolección y manejo de grandes cantidades de datos de registros. Permite de este modo que los usuarios puedan extraer el mayor provecho de los datos.

Es robusto, tolerante a fallos y distribuido, además su arquitectura es sencilla y flexible.

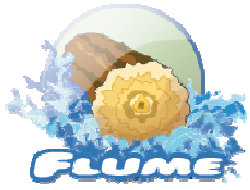


Ilustración 33. Logotipo Flume

3.9.17. WHIRR

Conjunto de librerías ofrecidas por Apache Software Foundation.

Ofrece un API general para ejecutar los servicios en la nube independientemente del proveedor, procurando una capa de abstracción en Java.

3.9.18. COUCH DB

Proyecto de código abierto creado por Apache Software Foundation.

Trabaja como una base de datos NoSQL con documentos de formato JSON. Tiene la ventaja de ser fácilmente escalable y fácilmente manejable. Hace uso de una interfaz REST[17] sobre protocolo HTTP.

Ilustración 34. Logotipo CouchDB

[17] Representational State Transfer, técnica de arquitectura software, formadas por un conjunto de convenciones que deben cumplirse.

3.9.19. MONGO DB

Base de datos NoSQL creada por la compañía 10gen, de código abierto desarrollada en C++.

Esta base de datos permite almacenar datos en unas estructuras propias denominadas BSON, similares a las estructuras JSON. Además permite adaptarse perfectamente a las características típicas de Hadoop como replicación, sistema distribuido, etc.

Ilustración 35. Logotipo MongoDB

3.9.20. ELASTICSEARCH

Herramienta de código abierto basada en Apache Lucene, bajo licencia Apache.

Se trata de un motor de búsqueda casi en tiempo real de forma distribuida y desarrollado en Java.

Ilustración 36. Logotipo Elasticsearch

3.9.21. SCRIBE

Herramienta de código abierto desarrollada por Facebook bajo licencia Apache.

Se trata de un servidor fiable y robusto, con diseño escalable en un gran número de nodos. Concebido para la agregación de datos de registro streaming.

3.9.22. CASCADING

Plataforma de aplicaciones de desarrollo en Apache Hadoop con licencia Apache y desarrollada en Java.

Permite a los programadores desarrollar y probar su aplicación a nivel local para posteriormente, tras salvar una etapa de pruebas, pueda ser lanzada. También permite una fácil integración con otras herramientas del ecosistema Hadoop y abstraer en parte de la complejidad de Hadoop.

Ilustración 37. Logotipo Cascading

3.9.23. BIGTOP

Proyecto de código abierto de Apache Software Foundation.

Herramienta que proporciona la creación de paquetes de desarrollo para interoperar con herramientas del ecosistema Hadoop, además de facilitar de entornos de pruebas para los mismos.

Ilustración 38. Logotipo Bigtop

3.9.24. LUCENE

Herramienta creada por Apache y escrita en Java.

Diseñada para la indexación y búsqueda de texto, independientemente del formato de fichero.

Además de la versión original se han desarrollado otras que permiten su uso con distintos lenguajes de programación, versiones también desarrolladas por Apache como Montezuma, Mutis, NLucene, PyLucene, Lupy, CLucene, PLucene, Lucene.NET… y otros subproyectos que mejoran el original como Lucene Core, Open Relevance Projecty Solr.

Ilustración 39. Logotipo Lucene

3.9.25. GIS TOOLS

Proyecto de código abierto creado por Esri bajo licencia Apache.

Se trata de un conjunto de herramientas que amplían las funcionalidades de Hadoop mediante una serie de librerías en el campo de la información geográfica.

3.9.26. HCATALOG

Proyecto de Apache Software Foundation.

Es una capa de gestión y almacenamiento de datos para Hadoop, se encarga de mejorar la lectura y escritura de datos en la red abstrayendo al usuario de la arquitectura usada.

3.10. PROYECTOS EN OTROS SISTEMAS

La compañía Hortonworks lanzó al mercado el proyecto Hortonworks Data Plataform, HDP, con el que se completa en gran medida el uso de la tecnología Hadoop en otros sistemas operativos como Windows.

Microsoft, quien ha trabajado de la mano de Hortonworks, también ha lanzado a través de Microsoft Azure su propia versión de la plataforma Hadoop para Windows, HDInsight.

3.11. SOLUCIONES FINALES

Anteriormente se ha visto como algunas empresas disponen del software y hardware necesario para que otras empresas puedan hacer uso de Hadoop en la nube, en este caso se va un paso más allá.

Para facilitar el uso de Hadoop a los usuarios son muchas las empresas que han desarrollado sus propias versiones de Hadoop ofreciendo soluciones finales al cliente.

Con estas plataformas las empresas son capaces de ofrecer un trato personalizado al cliente que cubra sus necesidades, además de ahorrarles la adquisición de las máquinas, el personal cualificado y demás. Mediante estas plataformas Hadoop está al alcance de todos independientemente del espacio o el personal con el que dispongan.

Estas empresas no ofrecen únicamente el servicio de Hadoop en la nube y su propia versión con mejoras y nuevas características importantes, sino que muchas de ellas además ofrecen un servicio integral de su uso, por lo tanto no tienen porque contar con una persona encargada de manejar este servicio ni en la nube ni en sus propias

instalaciones, simplemente es una empresa externa la que se encarga de realizar hasta el más mínimo detalle relacionado con Hadoop.

Algunas de las empresas que actualmente ofrecen estos servicios basados en Hadoop son:

- Cloudera: empresa de software estadounidense que ofrece servicios a empresas y particulares basados en Hadoop. Se diferencia en la creación de un software propio que mejora las características y servicios de Hadoop.

Su propia plataforma es conocida como Impala, mediante ella permite una notable mejora en la velocidad al realizar consultas de tipo SQL.

Ilustración 40. Logotipo Cloudera Impala

Además Cloudera ha lanzado al mercado Cloudera Search. Está basada también en Hadoop y se trata de una herramienta software que permite una búsqueda, manejo y análisis más sencillo y completo de los datos. Al igual que con Impala, Cloudera ofrece un software centrado en Hadoop con el que mejorar características y servicios básicos de Hadoop.

Ilustración 41. Logotipo Cloudera Search

- Oracle: compañía de software estadounidense fundada en 1977. Actualmente una de las más importantes compañías y con mayor crecimiento. Oracle lanzó al mercado "Oracle Big Data Appliance" en el año 2011 introduciendo con ello su propia versión de la plataforma Hadoop. Con ella ofrece nuevas funcionalidades y características mejoradas de Hadoop. Con el paquete "Oracle Big Data Appliance" dota a Hadoop de una nueva arquitectura SQL y nuevas funcionalidades en cuanto a redimensionamiento y seguridad, entre otras muchas funcionalidades.

Ilustración 42. Logotipo Oracle

- Microsoft: la compañía estadounidense fundada en 1975, ha lanzado HDInsight. Esta es su propia versión de la plataforma Hadoop. Esta distribución ofrece ya conocidas herramientas para Hadoop, como Pig, Hive, Ambari y demás, estando todas ellas, en esta versión, adaptadas a su uso en Windows. Además está diseñada para trabajar con Windows Server y Microsoft SQL Server.

Ilustración 43. Logotipo Microsoft

- IBM: la apuesta realizada por IBM es su software Infosphere. Con ello pretende disminuir los tiempos de procesamiento, agilizar y facilitar su uso automatizando la mayor parte de las funcionalidades, mejorar la migración entre aplicaciones y demás, todo ello basado en Hadoop.

Ilustración 44. Logotipo IBM

3.12. INSTALACIÓN

La instalación y ejecución de la plataforma se ha realizado sobre una máquina virtual, concretamente VmWare Player[18] 3.1.6 (Aunque existen versiones mucho más actuales, ha sido necesario utilizar una versión antigua por cumplir requisitos de software y hardware necesarios para su funcionamiento desde el PC utilizado para la realización del PFC) a la que se le ha realizado la instalación del sistema operativo Ubuntu[19] 32bits 12.04.4[20].

3.12.1. CONFIGURACIÓN STANDALONE

Esta configuración queda por defecto al realizar la instalación de Hadoop[21].

3.12.2. CONFIGURACIÓN PSEUDO-DISTRIBUIDA

Todos los ficheros nombrados a continuación para la configuración de Hadoop, pueden encontrarse bajo el directorio etc:

- Modificación core-site.xml:

```
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost>8020</value>
<description>Nombre sistema de ficheros</description>
</property>
</configuration>
```

Ilustración 45. Modificación archivo core-site.xml

[18] Puede encontrarse más información sobre este software en el glosario.
[19] Puede encontrarse más información sobre este sistema operativo en el glosario.
[20] Puede consultarse el anexo para más información y detalles de instalación.
[21] Puede consultarse el apéndice para obtener información a cerca de la instalación de Hadoop.

- Modificación hdfs-site.xml:

```
<configuration>
        <property>
                <name>dfs.namenode.name.dir</name>
                <value>file:/home/hadoop/workspace/dfs/name</value>
                <description>Ruta donde se almacenan los metadatos</description>
        </property>
        <property>
                <name>dfs.datanode.data.dir</name>
                <value>file:/home/hadoop/workspace/dfs/data</value>
                <description>Ruta donde el datanode almacena los bloques</description>
        </property>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
                <description>solo tenemos un nodo por tanto el factor de replicacion es 1</description>
        </property>
</configuration>
```

Ilustración 46. Modificación archivo hdfs-site.xml

- Modificación mapReduce-site.xml:

```
<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
        <property>
                <name>mapred.system.dir</name>
                <value>file:/home/hadoop/workspace/mapred/system</value>
                <final>true</final>
        </property>
        <property>
                <name>mapred.local.dir</name>
                <value>file:/home/hadoop/workspace/mapred/local</value>
                <final>true</final>
        </property>
</configuration>
```

Ilustración 47. Modificación archivo mapReduce-site.xml

- Modificación yarn-site.xml:

```
<configuration>
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
        <property>
                <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
                <value>org.apache.hadoop.mapred.ShuffleHandler</value>
        </property>
</configuration>
```

Ilustración 48. Modificación archivo yarn-site.xml

- Inicio de la plataforma Hadoop:

```
noelia@ubuntu:/usr/local/hadoop/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Incorrect configuration: namenode address dfs.namenode.servicerpc-address or dfs
.namenode.rpc-address is not configured.
Starting namenodes on []
noelia@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-noelia-na
menode-ubuntu.out
noelia@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-noelia-da
tanode-ubuntu.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is 1e:28:78:1d:3d:ab:12:1d:74:7c:cb:c0:8e:dd:f9:79.
Are you sure you want to continue connecting (yes/no)? y
Please type 'yes' or 'no': yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts
.
noelia@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-no
elia-secondarynamenode-ubuntu.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-noelia-resource
manager-ubuntu.out
noelia@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-noelia-n
odemanager-ubuntu.out
```

Ilustración 49. Comando para iniciar Hadoop tras su instalación

- Comando jps, muestra los procesos Java actualmente en proceso en el sistema:

```
noelia@ubuntu:/usr/local/hadoop/sbin$ jps
2869 ResourceManager
3193 Jps
2992 NodeManager
```

Ilustración 50. Procesos java en funcionamiento

- Interfaz gráfica para comprobar y hacer mediciones sobre el funcionamiento de los nodos:

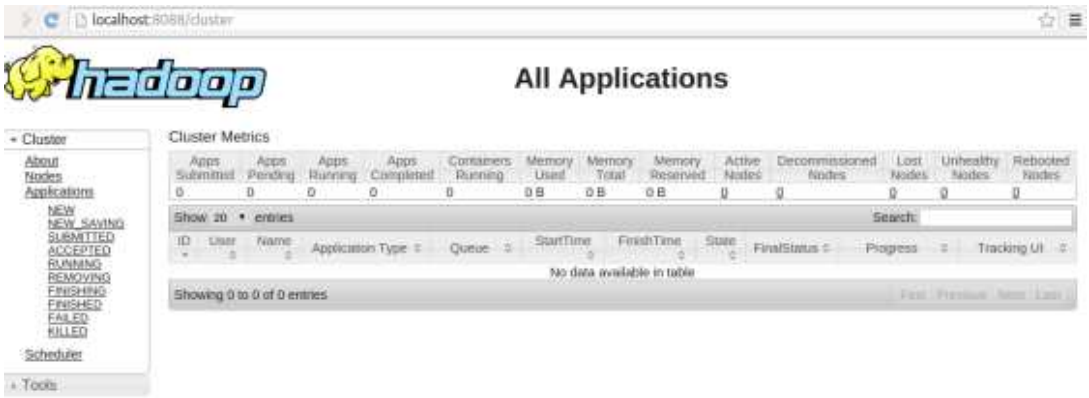


Ilustración 51. Interfaz gráfica sobre el funcionamiento de los nodos

Con esta configuración de Hadoop se puede comenzar a realizar los diferentes análisis que se desee, además se puede comprobar el rendimiento de los nodos en la interfaz gráfica durante el procesamiento.

4. CONCLUSIONES

Al realizar este proyecto me he dado cuenta de lo amplio que es este campo y de lo que aún queda por hacer.

Esto es sólo un aperitivo de lo que queda por llegar y que espero que sea la puerta a muchos otros estudiantes que utilicen este proyecto, en el que encontrarán toda la información comprimida y en un solo libro, para que les sirva como guía para desarrollar nuevas ideas.

También he podido descubrir la utilidad de los datos para predecir catástrofes, situaciones de riesgo y emergencias. Datos que antes eran inservibles o que únicamente se podían utilizar para realizar vagas estadísticas que no arrojaban previsión sobre ningún aspecto. Y es que la clave de de este campo está en la previsión, con Big Data y con herramientas como Hadoop se está abriendo un mundo de oportunidades al predecir todo aquello que se desea y de situaciones que ni siquiera se podía imaginar. Con esta herramienta no solo se abaratan costes y se generan mayores ingresos, sino que permite estar allí donde se deba antes de que algo ocurra.

5. BIBLIOGRAFÍA

Victor Mayer-Schönberger; Kenneth Cukier. *Big Data, la revolución de los datos masivos.* Madrid: Turner Publicaciones, 2013.

Jornada: El impacto de la Nube y el Big Data en la Ciencia, 21 de Marzo de 2013, Fundación Ramón Areces.

Thomas H. Davenport; Jill Dyché. *Big Data in Big Companies.* International Institute for Analytics, Mayo 2013.

Irwin King; Michael R. Lyu; Haiqin Yang. *Online Learning for Big Data Analytics.* The Chinese University of Hong Kong, Santa Clara, California, 2013.

James Manyika; Michael Chui; Brad Brown; Jacques Bughin; Richard Dobbs; Charles Roxburgh; Angela Hung Byers. *Big Data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute, May, 2011

Javier Alonso Cerrato; Igor García Olaizola. *Análisis Visual en el campo del Big Data: Visual Analytics.* Bit, junio 2013.

http://www.teradata.de/

http://big-project.eu/

www.computerwoche.de

http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html?cmp=BS&ct=SocialMedia&cr=twitter

http://bigdatauniversity.com

http://adayinbigdata.com/

http://cci.drexel.edu/bigdata/bigdata2013/index.htm

http://www.bigdataspain.org/

http://mexico.emc.com/campaign/bigdata/index.htm

http://es.wikipedia.org/wiki/Hadoop

http://es.wikipedia.org/wiki/Big_data

http://momentotic.wordpress.com/2013/05/16/que-es-hadoop/

http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/

http://blogthinkbig.com/hadoop-open-source-big-data/

http://www.ibm.com/developerworks/ssa/data/library/techarticle/dm-1209hadoopbigdata/

http://www.marketingdirecto.com/actualidad/bases-de-datos-y-crm/hadoop-el-software-que-ha-cambiado-la-vida-del-big-data/

http://www.cisco.com/web/ES/about/press/2013/index.html
http://www.dataversity.net

http://www.zdnet.com/

http://strata.oreilly.com/

http://www.datanalytics.com/

http://www.wired.com/

http://www.adictosaltrabajo.com/

http://www.happyminds.es/

http://www.javaworld.com/

http://aosabook.org/en/index.html

6. GLOSARIO

6.1. VMWARE PLAYER

Software gratuito creado por la compañía VMware fundada en 1998 en Estados Unidos.

Se trata de un software de virtualización, simula un sistema físico con unas características elegibles.

Este sistema de virtualización permite montar un sistema operativo virtual sobre una máquina física que no tiene porque funcionar con el mismo sistema operativo. La mayoría de los parámetros de la máquina virtual pueden ser elegidos por el usuario dependiendo de sus necesidades. Además las acciones que se realicen en la máquina virtual no tienen acción directa sobre la máquina física.

6.2. UBUNTU

Sistema operativo basado en Linux de código abierto lanzado en 2004 por la Fundación Ubuntu bajo licencia de software libre.

Todo el software y aplicaciones por las que está compuesto forman parte también de la familia de licencia libre a excepción algún controlador.

Actualmente este sistema operativo puede encontrarse además de en ordenadores de sobremesa y portátiles en una gran variedad de dispositivos como móviles, televisiones, tablets y demás. Actualmente se encuentra en continuo crecimiento y desarrollo. Sus actualizaciones se producen cada 6 meses además de otras diferentes actualizaciones en sus diferentes interfaces.

6.3. OPENSSH

Conjunto de aplicaciones desarrolladas por la empresa OpenBSD. Estas aplicaciones permiten realizar comunicaciones seguras bajo el protocolo SSH mediante la creación de túneles seguros y autenticación.

Debido a su naturaleza de comunicación entre máquinas está adaptado a trabajar bajo diferentes sistemas operativos sin problemas de manera ajena al usuario.

6.4. CREATIVE COMMONS

Organización sin ánimo de lucro fundada en el 2001. Permite usar y compartir software amparado bajo unos determinados instrumentos jurídicos. Aun manteniendo los derechos de autor permiten modificaciones del mismo para satisfacer el uso personal de cada usuario. Además también se han creado diferentes licencias creativas que amparan la voluntad del creador, pudiendo elegir así la que más le convenga.

Actualmente se encuentra aun en expansión mediante el proyecto Creative Commons International, que pretende hacer llegar este tipo de licencias a cualquier rincón del mundo.

Ilustración 52. Logotipo Creative Commons

6.5. BASES DE DATOS RELACIONALES

Esquema que cumple con el modelo relacional. Este permite establecer conexiones entre los datos almacenados en las diferentes columnas.

Actualmente es el tipo de base de datos más utilizada y son muchas las empresas y sistemas que permiten su uso, como Oracle, MySQL y Microsoft SQL Service entre otros. Para su uso es necesario el uso del lenguaje SQL (Structured Query Language) con el que interpretar cálculo y algebra lineal para realizar consultas, modificar y manejar las bases de datos, para obtener de ellas información de forma sencilla.

6.6. CORRELACIÓN Y CAUSALIDAD

Se dice que dos variables están correladas cuando la variación en una provoca un cambio en la otra.

En el procesamiento de las variables, pueden darse distintos tipos de causalidad:

- Causalidad directa: aquella fácilmente identificable en la que una variable es causante de los cambios sufridos en otra de manera inequívoca.
- Causalidad indirecta: la más complicada de verificar en la que existe una tercera variable que se relaciona indirectamente con otras variables.
- Causalidad con los datos: aquella que por el uso de la razón o la lógica son fácilmente descartables por ser causalidad sin sentido.

Por lo tanto se puede afirmar que la correlación no justifica la existencia de una relación causa y efecto.

6.7. GUARDIUM INPHOSPHERE

Software diseñado especialmente para el uso de Big Data bajo condiciones estrictas de seguridad por la compañía IBM.

Esta gama de productos promete proteger la seguridad, integridad y privacidad de la información, mediante diferentes funcionalidades como cifrado y descifrado de datos sin producir un gran impacto en latencia y procesamiento.

Además posee herramientas para localizar datos sensibles de ser publicados. Tiene la capacidad de detectar y eliminar datos de tipo privado que puedan ser compartidos. Trabaja a una gran velocidad, de forma prácticamente automática y en varios idiomas.

7. APÉNDICE

7.1. INSTALACIÓN DETALLADA PLATAFORMA HADOOP

Para la instalación y el funcionamiento de Hadoop se necesita de unas instalaciones previas.

Para el funcionamiento de Hadoop en esta demostración se requiere de la instalación de una maquina virtual, en este caso creada con VMWare Player, sobre la que instalar el sistema operativo Ubuntu.

En el sistema operativo Ubuntu es necesario desactivar IPv6 como la normativa de Hadoop indica. Además se requiere de la instalación de Java y de la instalación y configuración de ssh.

Tras ello es posible comenzar la instalación de Hadoop, que se realiza de forma sencilla y rápida una vez se descarga.

Tras la instalación, Hadoop queda configurado por defecto en modo Standalone[22]

A lo largo de los siguientes apartados puede obtenerse información detallada sobre cada una de las instalaciones de software necesarias para instalar Hadoop.

7.1.1. INSTALACIÓN VMWARE PLAYER

Se descarga VMWare Player:

https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/3_0|PLAYER-316|product_downloads

Una vez descargada la versión 3.1 de VMWare Player se comienza con su instalación.

[22] Pueden consultarse otras configuraciones en el apartado 3.12.2

1. Se selecciona siguiente (Next>)

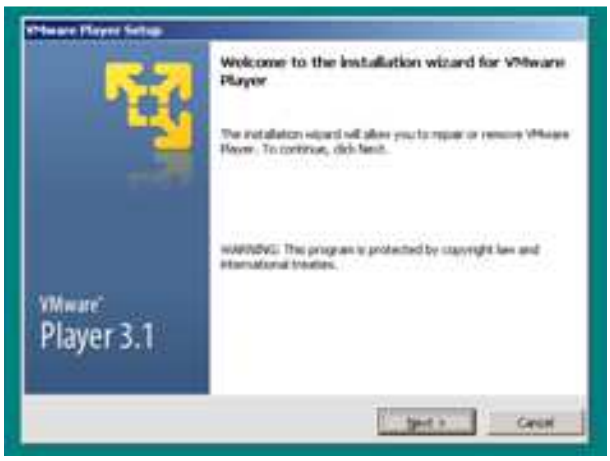


Ilustración 53. Inicio instalación VMWare Player

2. Se elige la ubicación donde se desee instalar el software y se selecciona siguiente (Next>)

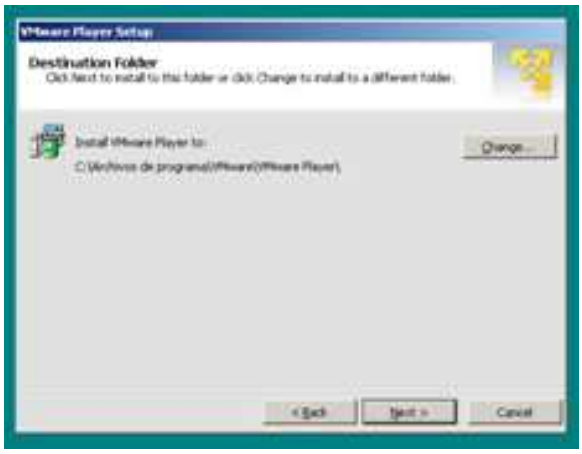


Ilustración 54. Elección carpeta instalación VMWare Player

3. Se selecciona si se quiere obtener las actualizaciones del software

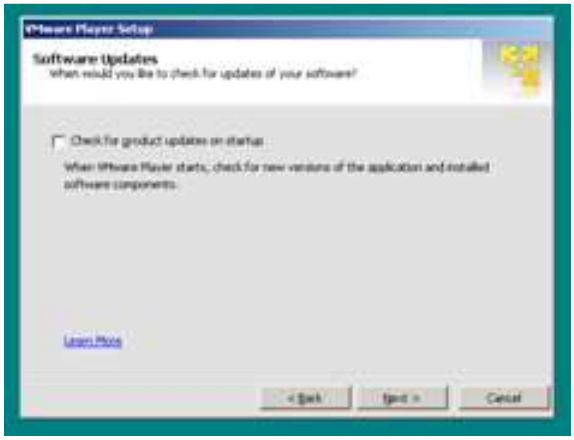


Ilustración 55. Actualizaciones VMWare Player

4. Se selecciona si se desea enviar información anónima para la comprobación de funcionamiento, errores y estudios estadísticos

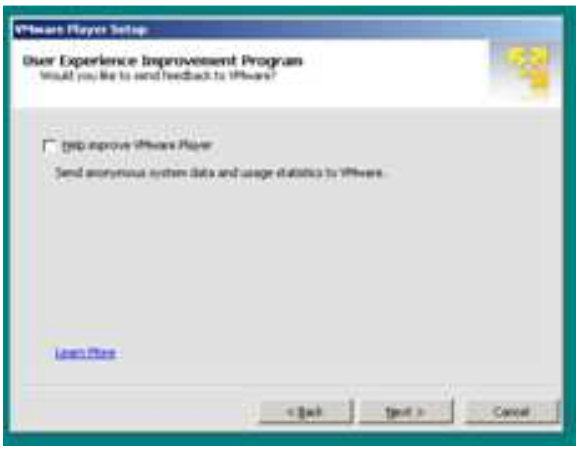


Ilustración 56. Información para supervisión VMWare Player

5. Por último antes de que comience la instalación se elige si se quiere lanzar directamente el programa tras su instalación o si se desea hacer accesos directos al mismo en el escritorio o en la pestaña de programas

Ilustración 57. Modo inicio VMWare Player

6. Para comenzar la instalación después de haber elegido todas las opciones sobre su instalación, se pulsa el botón continuar (Continue)

Ilustración 58. Finalización instalación VMWare Player

7. Se puede comprobar el estado de la instalación mediante el progreso en la barra y una vez a concluido se pulsa el botón siguiente (Next>) para finalizar

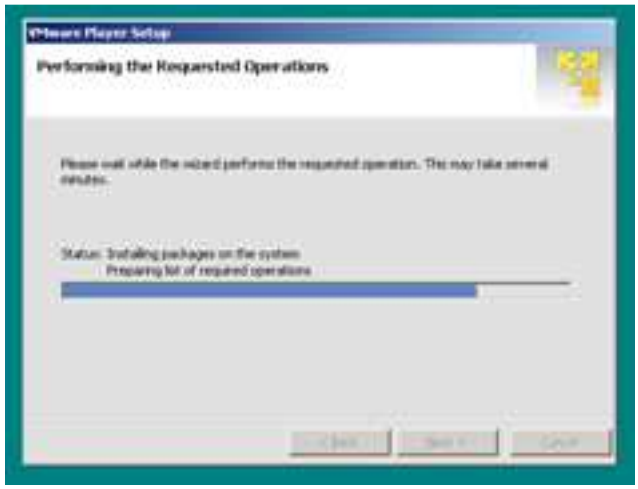


Ilustración 59. Proceso instalación VMWare Player

8. Al concluir la instalación se ofrece la oportunidad de reiniciar el equipo en el mismo momento o reiniciarlo más tarde

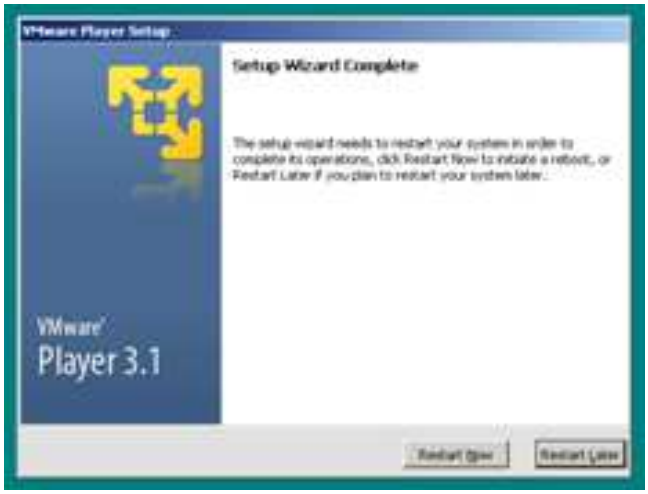


Ilustración 60. Reinicio equipo tras completar instalación VMWare Player

9. Una vez el programa está instalado y el equipo reiniciado se puede acceder al software para crear la máquina virtual e instalarle un sistema operativo sobre el que luego se trabajará. Para comenzar creando la máquina virtual se selecciona la opción “Crear una nueva máquina virtual” (Create a New Virtual Machine)

Ilustración 61. Ventana inicio VMWare Player

7.1.2. INSTALACIÓN UBUNTU

Se descarga el sistema operativo Ubuntu, se elige la versión teniendo en cuenta las capacidades del sistema en el que va a ser instalado:

http://www.ubuntu.com/download/desktop

1. Una vez se ha seleccionado la opción de “Crear una máquina virtual” y se dispone del sistema operativo descargado en el equipo, se va a seleccionar la ubicación donde se encuentra almacenado para proceder con las características básicas que tendrá la máquina virtual y el sistema operativo. Una vez se tenga localizado el sistema operativo elegido se pulsa siguiente (Next>)

Ilustración 62. Ventana VMWare Player para elección del sistema operativo

2. Se completan las credenciales contraseña y nombre de usuario

Ilustración 63. Configuración de credenciales para sistema operativo en VMWare Player

3. Se elige el nombre que tendrá la máquina y la ubicación donde se almacenara el equipo real

Ilustración 64. Configuración nombre de la maquina en VMWare Player

4. Se muestra la ventana en la que hacer las elecciones de espacio de disco duro

Ilustración 65. Configuración parámetros en VMWare Player

5. Tras elegir el tamaño del disco duro se muestra la ventana donde configurar los parámetros que tendrá el sistema, se pueden dejar por defecto o configurarlos manualmente seleccionando “Customize Hardware”

Ilustración 66. Configuración parámetros avanzados en VMWare Player

6. Al pulsar en la pantalla anterior “Finish” se comienza con la instalación de sistema operativo

Ilustración 67. Inicio del sistema operativo en VMWare Player

7. Comienza la carga del sistema operativo

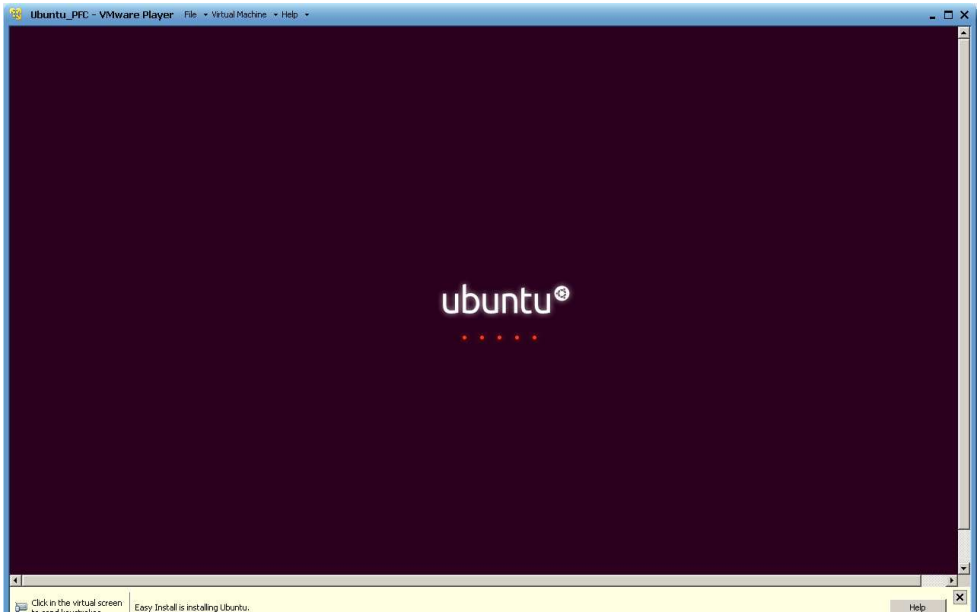


Ilustración 68. Comienza la instalación del sistema operativo sobre VMWare Player

8. Una vez la máquina virtual ha arrancado el sistema operativo se muestra el proceso de instalación

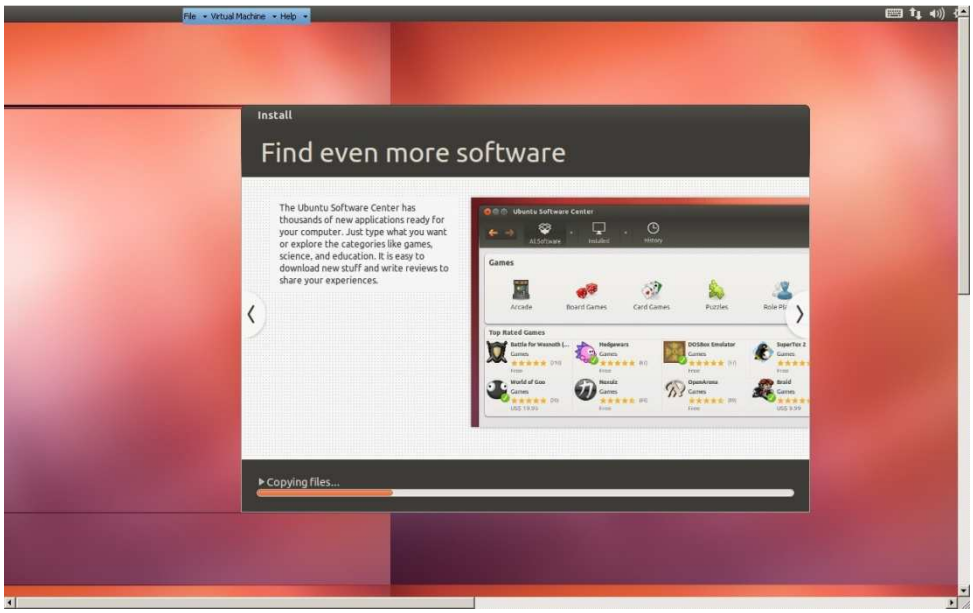


Ilustración 69. Proceso de instalación del sistema operativo sobre VMWare Player

9. Tras la completa instalación se inicia el sistema operativo (en el caso de este ejemplo se inicia en ventana de comandos)

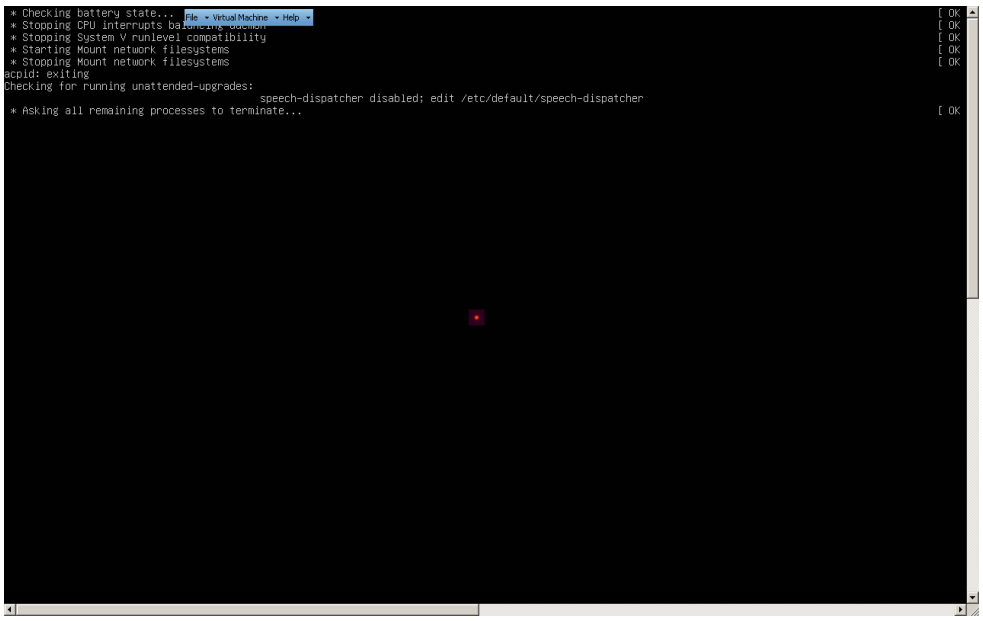


Ilustración 70. Sistema operativo funcionando sobre VMWare Player

7.1.3. INSTALACIÓN JAVA 1.6

1. Una vez el sistema operativo está instalado, se comienza la instalación del software necesario, para ello se accede al sistema mediante el login

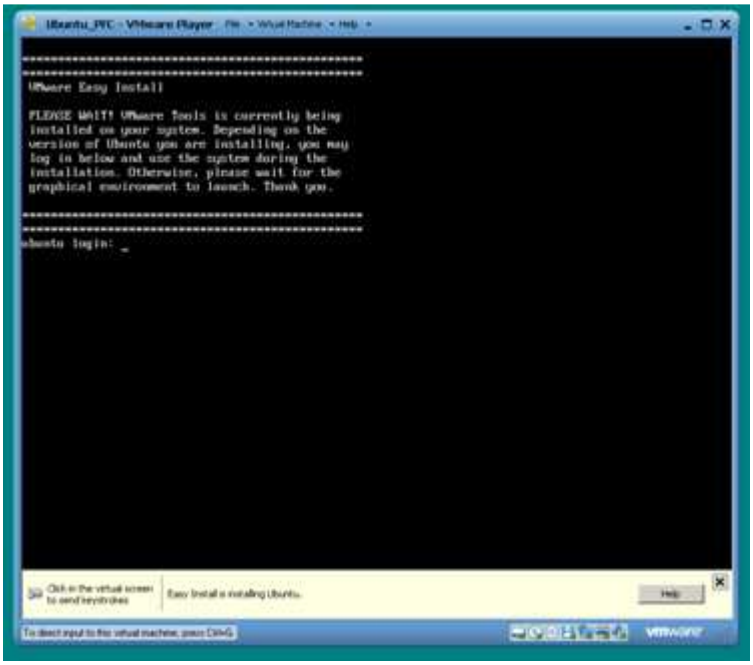


Ilustración 71. Introducción de usuario y contraseña en el sistema

2. Tras estar identificado en el sistema correctamente se puede comenzar la instalación de Java escribiendo los comandos en el prompt

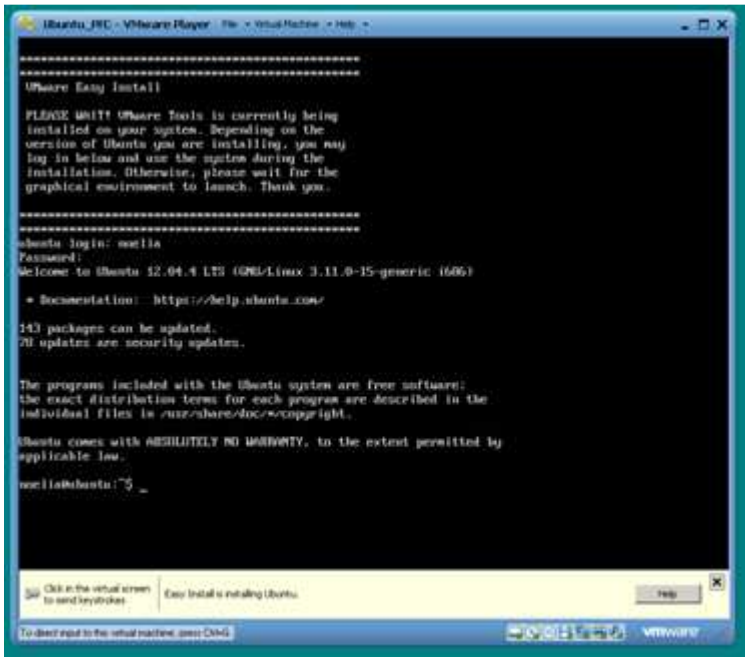


Ilustración 72. Sistema operativo tras ser correctamente identificado el usuario

3. Se descarga la versión de java deseada

```
noelia@ubuntu:~$ sudo add-apt-repository ppa:webupd8team/java
 Oracle Java (JDK) Installer (automatically downloads and installs Oracle JDK6 / JDK7 / JDK8). Th
 are no actual Java files in this PPA.

More info:
- for Oracle Java 7: http://www.webupd8.org/2012/01/install-oracle-java-jdk-7-in-ubuntu-via.html
- for Oracle Java 8: http://www.webupd8.org/2012/09/install-oracle-java-8-in-ubuntu-via-ppa.html

Debian installation instructions: http://www.webupd8.org/2012/06/how-to-install-oracle-java-7-in-
ian.html
 More info: https://launchpad.net/~webupd8team/+archive/ubuntu/java
Press [ENTER] to continue or ctrl-c to cancel adding it
```

Ilustración 73. Descarga de Java desde línea de comandos

4. Una vez descargada se procede a la instalación mediante los comandos correspondientes

```
noelia@ubuntu:~$ sudo apt-get install oracle-java7-installer
Reading package lists... Done
Building dependency tree
Reading state information... Done
oracle-java7-installer is already the newest version.
0 upgraded, 0 newly installed, 0 to remove and 364 not upgraded.
```

Ilustración 74. Instalación de Java desde línea de comandos

5. Por último se comprueba que ha sido correctamente instalada a la vez que se comprueba la versión

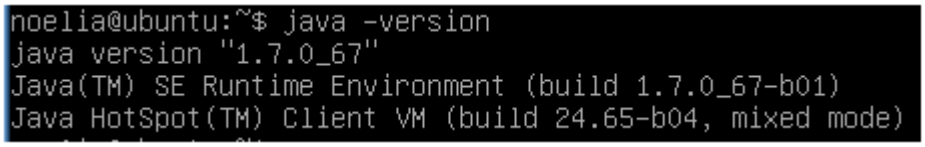


Ilustración 75. Comprobación de la versión de Java instalada

7.1.4. CONFIGURACIÓN SSH

1. Se procede a la instalación de ssh[23] mediante los comandos seguidos al prompt

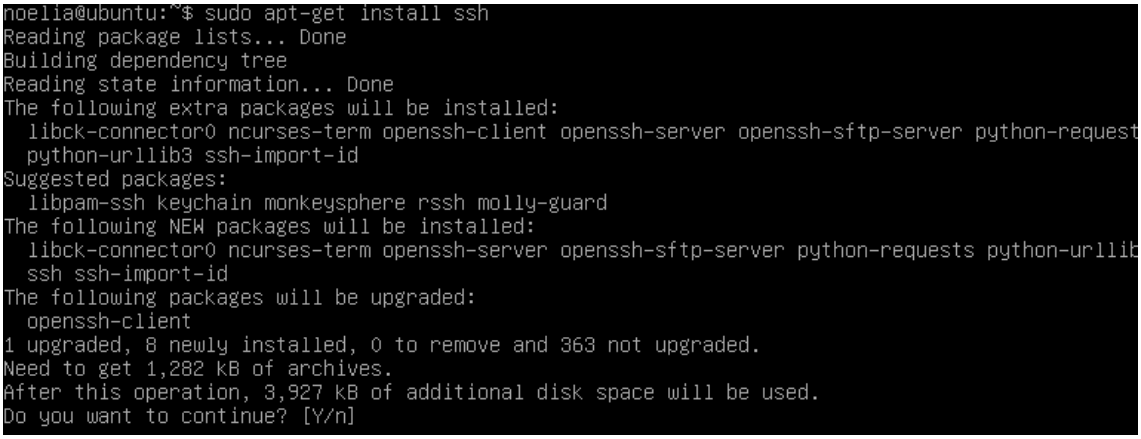


Ilustración 76. Instalación de ssh en el sistema operativo

2. Una vez se ha terminado la instalación, se comprueba que se tenga acceso al servidor mediante el comando

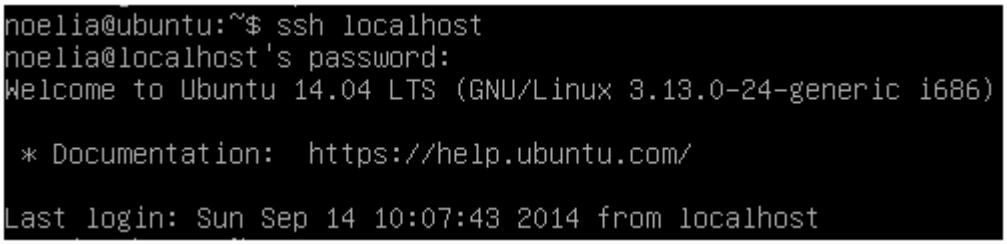


Ilustración 77. Acceso al host de la máquina

[23] Puede encontrarse más información acerca de este protocolo en el glosario.

7.1.5. CONFIGURACIÓN IP VERSIÓN 6

1. Se desactiva IPv6 como recomienda la normativa de Hadoop. Se comprueba el estado. La devolución de un 0 significa que IPv6 está activado.

```
noelia@ubuntu:/etc$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
0
```

Ilustración 78. Comprobación estado de IPv6

2. Para desactivar IPv6 se modifica el archivo sysctl.conf que se encuentra bajo /etc

```
noelia@ubuntu:/usr/local/hadoop/etc/hadoop$ ls
capacity-scheduler.xml      hdfs-site.xml              mapred-site.xml.template
configuration.xsl           httpfs-env.sh              slaves
container-executor.cfg      httpfs-log4j.properties    ssl-client.xml.example
core-site.xml               httpfs-signature.secret    ssl-server.xml.example
hadoop-env.cmd              httpfs-site.xml            yarn-env.cmd
hadoop-env.sh               log4j.properties           yarn-env.sh
hadoop-metrics2.properties  mapred-env.cmd             yarn-site.xml
hadoop-metrics.properties   mapred-env.sh
hadoop-policy.xml           mapred-queues.xml.template
noelia@ubuntu:/usr/local/hadoop/etc/hadoop$
```

Ilustración 79. Localización de ficheros bajo /etc

Y se añade al archivo las líneas:

```
#net.ipv6.conf.all.disable_ipv6=1
net.ipv6.conf.default.disable_ipv6=1
net.ipv6.conf.lo.disable_ipv6=1
```

Ilustración 80. Líneas a añadir al fichero para desactivar IPv6

3. Se actualiza

```
noelia@ubuntu:/etc$ sudo sysctl -p
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

Ilustración 81. Actualización fichero para desactivación de IPv6

4. Y por último se comprueba si ipv6 se ha desactivado

```
noelia@ubuntu:/etc$ cat /proc/sys/net/ipv6/conf/all/disable_ipv6
1
```

Ilustración 82. Comprobación estado de IPv6

7.1.6. INSTALACIÓN HADOOP

Primero se descarga Hadoop, en esta demostración se ha descargado la versión 2.2.0.

http://mirror.dkd.de/apache/hadoop/common/hadoop-2.2.0/

1. Se descomprime e instala el archivo

```
noelia@ubuntu:~/Downloads$ tar xzf hadoop-2.2.0.tar.gz
```

Ilustración 83. Descompresión e instalación de Hadoop

2. Se cambia la ubicación de la instalación de Hadoop para mayor comodidad posteriormente

```
noelia@ubuntu:~/Downloads$ sudo mv hadoop-2.2.0 /usr/local/
noelia@ubuntu:~/Downloads$ sudo mv /usr/local/hadoop-2.2.0 /usr/local/hadoop
```

Ilustración 84. Cambio de ubicación para la instalación de Hadoop en el sistema

3. Se añade al fichero hadoop-env.sh la ruta donde se encuentra Java instalado, en esta demostración se encuentra bajo la ruta:

export JAVA_HOME=/usr/lib/jvm/java-7-oracle

```
noelia@ubuntu:/usr/lib$ cd jvm/java-7-oracle/
bin/     db/      include/ jre/     lib/     man/
```

Ilustración 85. Actualización de la ruta donde se encuentra Java

7.2. API MAPREDUCE

- Versión actualizada del API MapReduce: http://hadoop.apache.org/docs/r2.3.0/api/org/apache/hadoop/mapReduce/package-summary.html

- API MapReduce: http://hadoop.apache.org/docs/r1.1.1/api/org/apache/hadoop/mapred/package-summary.html

7.3. API HDFS

http://hadoop.apache.org/docs/current/api/org/apache/hadoop/fs/FileSystem.html

7.4. API HADOOP

http://hadoop.apache.org/docs/current/api/

7.5. DIRECCIONES INTERESANTES

- Es posible gracias a Big Data

  http://worldwind.arc.nasa.gov/java/

  https://www.google.com/sky/

- Base de datos de libre acceso del gobierno de España

  http://datos.gob.es/

- Base de datos de libre acceso del gobierno Alemán

  https://www.govdata.de/

- Base de datos Estadounidense OSC

  www.opensource.gov

- Base de datos australiana NOSIC

  http://www.nosic.com.au/

7.6. COMPAÑIAS QUE HACEN USO DE BIG DATA

Algunas de las compañías que hacen uso de Big Data durante el desarrollo habitual de su trabajo son:

1&1
A9.com
AM Biotech
Amazon
American Apparel
Ancestry.com
AOL
Apixio
Beach Mint
Booz Allen Hamilton
CastLight
Chevron

China Telecom
Codeacademy
Cosmos Bank
DHL
DPR Construction
eBay
EHarmony
El California ISO
Facebook
Ford
Foursquare
Fox Interactive Media
Freebase
General Electrics
Hydro One Networks
IBM
ImageShack
IRS Compliance Data Warehouse
ISI
Joost
Last.fm
LinkedIn
Meebo
Metaweb
Mitula15
Mount Sinai Medical Center
NARA ERA
NASA Human SpaceflightImagery
Ning
NOAA NWS
Nokia
Oklahoma Gas & Electric
Oncor
Pepco
Powerset
Rackspace
Salesforce
San Diego Gas y Electric
Seton
StumbleUpon16
Telecom
Telefónica
TerraEchos
The New York Times
Tuenti

Twitter
Veoh
Vesta Wind Energy
Volvo
Wallmarkt
Zoosk